
Co-China周刊
NO.156
2013年5月30日
可疑的大数据
月末版“失控”的新概念 第一期
我在中國 論壇
Co-China

编者的话

我们大可以这样过一天了。早上醒来打开手机用各种 APP 浏览全球动态，随后打开音乐软件选取那些一辈子也没听过的老爵士，看着扫地机器人来来回回清洁屋子，中午在线点餐吃一顿批萨，下午用网络电视看一会儿不知道什么国家的动物世界，傍晚带着智能手环跑个 3、5 公里，看部 4k 电影，临睡前处理来自大洋彼岸的邮件，刷刷微博、微信，各种点赞指点江山，睡着后智能手环还将持续记录你的心跳、评估睡眠，在最恰当的时候温柔地提醒你，又是新的一天了。

50 年，不，20 年前，这一切听起来都十分迷人；不用费心多幻想，未来 50 年，也许 20 年，这一切都将升级为一个更炫酷的未来。我们会有更智能的社交、交通、娱乐、教育方式，我们将会获得更全面的趋势分析，进入更深的海域，突破更远的太空，联通一些更聪明的替身，甚至真正找到长生不老的密钥，这些吸引着人类在“更快、更高、更强”的路上孜孜以求，发现、探索、创造，人类突破自然规律，将生物逻辑注入芯片、电子网络、机器人模块、药物搜索、企业管理等等之中，将自然与科技相融合铸就未来，而这个人类打造的未来会是怎样呢——走向迷人，抑或失控?

凯文 · 凯利在其著作《失控》中从人工智能、生态学、仿真学、机器人技术、网络科技等诸多领域入手，为全人类的最终命运和结局做了结语。 当然，结局并非如该书标题这样简单粗暴。任何一种新技术的发展都为人类开启一扇从未开启的窗，然而同时是否在某处悄悄关上了一扇窗，如同开什么窗般未可知。人类的结局这个题目太大，对历史长河中渺小的我们而言，新概念、新技术为我们带来了无穷的可能性，伴随着这些可能性问题和矛盾也

从生。在探究过人类之间的关系之后，周刊将开启关于新概念与新技术的系列月末刊，为身在科技高速发展世界中的我们，环顾四周那些正在蓬勃前进或受到人们热议的新概念（技术），尝试看看在这些迷人的未可知背后，是否还有值得我们了解的关于“失控”的未可知。

本系列的第一期我们从“大数据”开始。《大数据时代》一书的畅销，在全球挂起关于数据的飓风，大至国家安全小到球赛预测，随着云技术的发展大数据“一切皆有可能”的身影活跃在任何一个行业任何一个人的周围，然而大数据真有这么神吗？本期周刊，我们就将告诉大家大数据有哪些解决不了的问题，也从隐私、道德、文化等不同的角度，解读大数据可能引发的“失控”。

当然，在这里我们无意指摘、批评任何一种新概念，而是在如火如荼奔跑的途中，尝试给大家提供一种别样的风景，也许这是一条通向未来的岔路，也许那也不过是条死胡同。毕竟，未来的事，谁知道呢？

目录













【人人都玩大数据】

Joshua Brustein：大数据到底怎么影响我们的生活？



Joshua Brustein：《商业周刊》网站作者

“

大信息大爆炸的今天，不讨论大数据这个话题似乎就是跟不上时代。从医药到教育，再到其他各个领域，大数据充斥着现代社会的每个角落。而我们最关心的还是大数据最终将以什么样的形式，怎么样影响甚至改变我们的生活。

”

在信息大爆炸的今天，不讨论大数据这个话题似乎就是跟不上时代。从医药到教育，再到其他各个领域，大数据充斥着现代社会的每个角落。而我们最关心的还是大数据最终将以什么样的形式，怎么样影响甚至改变我们的生活。来听听四位专家告诉你大数据到底有多少可能。



丹•瓦格纳 Dan Wagner　　Civis Analytics 的创始人兼首席执行官



你曾经说过，希望用大数据解决全球最大的问题。你最想解决的问题是什么？

Dan Wagner：我们主要关注两个领域：教育和健康。在教育领域，我们专注于利用个人层面的数据，帮助客户发现那些申请和注册的大学低于其潜能的低收入学生，并帮助这些机构找到适当的方法，让这些孩子进入与其潜能相匹配的大学。

保险投保也是我们的目标之一，尤其是在《平价医保法案》刚开始施行的头几个月。我们主要致力于与多家机构合作，帮助它们找到没有医疗保险的民众，并让他们加入到医保计划中来。这是一项非常艰巨的工作，因为没有现成的无保险人员名册。你只能通过统计推断来完成这项工作。

最值得关注的问题之一，是保险如何从团体保险向个人保险发展，以及保险公司如何学会管理这一风险。我们正同几家机构合作，利用数据提前发现诸如心血管疾病等个体风险，提前了解病人面临的风险。

一旦发现有风险，你会增加投保人的保费吗？

Dan Wagner：你不能这么做。你只能根据诸如年龄等一系列精简变量来确定保单价值。因此，你不能根据上述风险来定价，但你需要管理这一风险。

你同奥巴马竞选团队合作时，大数据发挥了怎样的作用？

Dan Wagner：我们带来的最显著改变是在媒体方面。具体来说，就是利用尼尔森收视率来追踪竞选广告的投放和效果。透过收视率数据，你就好像看到了一张人口统计表，能从中了解到观众群的构成，例如是西班牙裔，还是女性。

我的做法是，根据我们计算得出的个人可说服得分来定义我们的观众。我们将这些数据与机顶盒数据相匹配。由此就能找到每一美元广告投放能带来最高可说服观众密度的电视栏目。有了这些数据，我们基本就能根据一个单一的标准来决定广告投放的优先顺序。这与人口统计学没有任何关系。只需明确哪些是我们在个人层面上确定的、要特别针对的观众群。这是一项非常艰巨的工作，但从文化角度来看，这种方法很适合我们的竞选团队，因为，奥巴马竞选的典型特征是，选民摇摆不定。

我们应当如何解决数据分析中的安全问题？

Dan Wagner：你必须非常重视这个问题。很多这类机构在收集信息，但我认为，其中很多机构都没有意识到什么是最高标准的数据安全操作。我们的很多工作都是在亚马逊云服务

平台上完成的，这比你内部可能开发的东西要好得多，因为你可以利用亚马逊提供的很多网络协议。亚马逊的云计算服务算是该领域最好的。

大数据热潮中，我们可能犯的最大错误，或可能忽视的最重要问题是什么？

Dan Wagner：大数据热潮最令人遗憾的一点是，人们只考虑其过程，而没有考虑结果。有些时候，这股热潮有些盲目，在某种意义上，它只是将对数据计算能力增长的信念孤立地看作是一种解决问题的手段。你将如何运用这些未来真的能改善人们生活的数据？这是个大问题。

在日常生活中，你是如何应对信息过载问题的？

Dan Wagner：作为一个在互联网相关公司工作的人，我有很多时间是在网上。但我尽量缩短通过各类电子设备进行沟通的时间，并努力确保自己阅读大量书籍。



埃里克•谢德特 Eric Schadt，伊坎基因组织学和多尺度生物学研究所(Icahn Institute for Genomics and Multiscale Biology)董事

如何证明超级计算在医学研究中能发挥重要作用？

Eric Schadt：主要通过两种途径。一是管理当下医学领域产生的诸如 DNA 测序等海量数据。举例来说，一位癌症病人的全基因组序列会产生万亿字节之多的数据。想象一下，如果要为数十万人做基因测序，就会产生千万亿，甚至百亿亿字节量级的数据。要对这些数据进行管理并加以处理，使之转化为能被医界人员所用的信息，就需要超级计算设备和相关的专业知识。

另一个途径是，利用需要超级计算在短时间内完成的非常复杂的数学算法，根据已经存在的疾病亚型，以及治疗该疾病亚型可能的最佳方法建立一个疾病预测模型。

这使医生在治疗中的作用以及病人与医生间的数据关系发生了怎样的变化？

Eric Schadt：发生了根本性的变化。与我们现有方法的不同之处在于，我们更深入地研究个体，而非一个群体。就拿糖尿病来说，目前可能有 100 种不同的糖尿病亚型，而且你和你的邻居得这种病的原因也各不相同。你可能是因胰腺 β 细胞功能受损所致；或者你肌肉中的摄取受体不能有效地吸收葡萄糖等等。不同的病因可能需要不同的治疗方法。

医生看到的只是晚期症状，但现在透过各类分辨率更高的科技产品他们能看到导致下游结果的上游病因。最近医生们才看到了这些病因。其中涉及数百万个变量，这是人脑无法理解的。

您刚才说到的都是数学帮助克服人脑缺陷的方面，这些数学计算程序有哪些缺陷需要人脑的帮助？

Eric Schadt：我们所做的工作是用一种需要人脑参与的方式来呈现信息，这是一种很棒的模式识别机器。目前在很大程度上人与机器是合作伙伴关系。也许 10 年、20 年以后，诸如 Watson 等计算机将变得非常强大，人的干预会大大降低。但目前还做不到。

很多组织收集的医学数据只供己用，我们应该对此感到担忧吗？

Eric Schadt：如果我们真的希望对人类健康产生影响，这些数据和模型必须对所有人进行开放。

物理研究领域就有这样的先例，强子对撞机试验的全部数据都是对公众开放的。当然，存在如何保护个人隐私的问题。

可以通过技术解决隐私保护问题吗？

Eric Schadt：我们当然可以保护并存储数据，保护计算机环境的安全，并采取众多安全协议来确保数据不会陷入危险。但有一点我们很清楚，任何形式的高维数据都无法真正做到匿名。就像照片一样。你不能指望你的外貌也有隐私，因为人人都能看到你的脸，你不能将它藏起来。我认为 DNA 以及诸如分子尺寸等其他数据最终也将归入同样的范畴，原因很简单，当技术足够成熟的时候，基因测序会像照相一样简单、便宜。

在日常生活中，你是如何应对信息过载问题的？

Eric Schadt：不能陷在大数据中。我会去玩单板滑雪、骑摩托车，或是做一些能帮助你放松，无需太动脑筋的活动。



安德烈斯•维根 Andreas Weigend，独立顾问，亚马逊公司前任首席科学家

你曾经将大数据比作原油。

如果你在后院发现了原油，你的这个发现可能用处不大，因为你需要将原油精炼后才能供人们使用。原始数据也像原油一样，不是拿来就可使用。亚马逊和谷歌就是从事数据精炼提取的公司。当然，据我所知，原油和数据两者之间最大的区别是，数据一时半会儿不会被用光。而至于价格，信息产品和石油产品之间的关系也非常有意思。

原油的大部分好处被你所描述的精炼公司而不是被其用户获得。我们怎样才能保证每个人都能从大数据中获益？

Eric Schadt：我认为，在苹果公司的应用商店发生的一切将会在数据经济领域再次上演，会有公司以这些数据为“原材料”推出服务。如果成立一个应用商店，而另一家公司使用数据向消费者提供应用并与数据公司共享收入，价值由此产生。

大约 10 年前，你曾担任亚马逊首席科学家。目前，世界是否已经完全变样了？

Eric Schadt：10 年前，我们已经注意到行业的重点正由算法（意味着你可以从自己所有的数据中获得任何东西）向仅仅需要获得更多的原始数据这一方向转移。所以说，现在的情况与当年完全不同，不过，我们仍然有类似的想法。贝索斯还是贝索斯。

你认为哪些行业守着最丰富的数据金矿，却未找到利用金矿创造价值的方法？

Eric Schadt：中国有一家公司名叫腾讯，他们推出的微信完全颠覆了中国人的沟通方式。与之相对应的另一家公司是中国电商公司阿里巴巴，该公司了解客户对哪些商品感兴趣，

他们在搜索什么商品，以及他们最终买了哪些商品。阿里还清楚客户是否退货和有无付款问题等。

这两家公司均拥有 10 亿客户。它们了解客户的所有沟通习惯或所有财务交易情况，所以，它们确实大有可为。当然，这也取决于你对哪些行业感兴趣。不过，真正的潜力是这两方面数据的交叉整合。比如，在需要做出信贷决定时，你可以从腾讯了解很多信息。因为，了解到你是否曾经和妓女鬼混或与拉皮条的家伙有过来往，也能多多少少地了解你将来偿还贷款的倾向。（老外对中国大数据的研究真透彻）

在日常生活中，你是如何应对信息过载这一问题的？

Eric Schadt：我们必须形成一个认知习惯，认识到人们是会错过一些信息的。如果有人错过了你的一封电子邮件，请不要生气。请通过另一个渠道与他们联络。



威廉•库科尔斯基 William Cukierski，Kaggle 公司的数据科学家

效果最好的竞赛有哪些？

William Cukiersk：我最看好的一场竞赛叫“找鲸大赛”。竞赛中要寻找的鲸是生活在大西洋中的一种濒危种群。这些搜寻者拥有强大的网络，不间断地记录鲸发出的声音，他们也拥有自己的算法，且效果非常好。他们说：“要不我们把这些数据交给 Kaggle，看 Kaggle 有没有更好的解决方案。”他们最后实现了非凡的成果。目前，这些强大的网络能够以接近 99% 的准确率来侦测出鲸的声音。我认为，如果有人坐在纽约的办公桌前就可以



从事与日常工作毫无相干且在万里之遥的一项工作，并为我们的日常生活带来巨大好处，这将是一项多么了不起的事情!

你们还在设法利用数据分析来进行癌症研究。Kaggle 是否组织过很多医疗相关领域的竞赛？

William Cukiersk：Kaggle 尚未在医疗领域涉足过多，主要原因是涉及泄露患者信息这个问题。另一个难题是拥有这些数据的个人和机构把数据囤积了起来，不愿分享。

制药公司拥有制药试验的数据，它们把这些数据压在了箱底。人们为了数据分享作了一些初步努力，也承诺在这方面展开合作，但结果还是各自都想保留自己手中的数据。从某种程度上说，主要还是担心隐私保护问题。你可能不会愿意把别人的基因组公开发布，然后大家都看出来这是家住主干道 232 号的萨利•斯密斯(Sally Smith)的基因组。不过，与此同时，这些担心也有些过度。对于这种问题，人们好像都在玩花招，说什么除非把数据直接交给你，不然你怎么能够远距离地利用数据解决问题呢？如果能消除这些顾虑，你就可以取得一些实质性的进展。

你们公司在举办人人都可以参与的竞赛，而有些占有数据的机构却牢牢抓着数据不放手。这是否是一个矛盾？

William Cukiersk：我在日常工作中面临的最大挑战之一是说服人们分享数据，并令其确信这么做不会威胁到其机构的生存。

经常情况下，不是说你占有了数据，数据就成为与生俱来的无价之宝，数据是需要挖掘和分析的。如果我们从一个机构拿到了一组数据，并将其公开，问题的解决方式是公开的，这不会产生什么问题，因为没有其他人有相同的数据，也没有人会再去获得并利用这些数据。

你认为，关于大数据的各种说法和观点，哪方面的失控最严重？

William Cukiersk：我必须纠正一下你的问题，应该是哪些方面没有失控。在与人们谈论大数据时，很难避免失控这个问题，也很难避免其老板的介入，同样难以获得老板支持地说“好吧，我们也做大数据吧”。我认为，人们在数据量方面有些失控。所以，经常有人会说，“我们有十亿兆的数据，我们有百万兆的数据。”许多问题可以在更小的数据规模上得到解决。比如，用输送带来筛选利马豆。销售利马豆的公司希望利用照相机来发现输送带上变质的利马豆。你可以想象，如果你能够发现一粒棕色利马豆，你就可以发现所有



的棕色利马豆，而不需天文级别的数据来解决这一问题。我认为，95% 的问题适用于这个模型。剩余 5% 的问题的算法需要大量的数据，提供的数据越多，其方案的效果就越好。Netflix 向用户推荐电影就是最好的例证。

（文章来源：36 大数据）

【原文链接】　　【回到目录】





大数据如何影响文化产品

近年来，大数据已经成为一个热门话题。它不仅影响着电子商务、金融服务、零售、医疗健康等行业，也开始在文化产业的各个领域发挥作用。

根据美国的统计资料，文化传媒行业数据是仅次于政府信息数据的第二大数据来源。文化产业本身就能够不断地产生或获得新的数据资源。如何有效利用这些庞大数据，提高对消费者需求和偏好的了解，生产出更符合目标人群的商品，实现价值提升，已成为文化企业的新课题。本期“环球参考”以新近案例呈现大数据在外国影视、音乐、旅游、游戏、设计等领域的具体应用。

影视：只拍观众想看的

过去，制作一部影视剧，主要靠导演、制片人的经验；投资一个影视项目，投资机构很多时候是依赖人脉关系网；播出一部影视剧，则主要借力播出平台的品牌效应。在这些环节中，观众作为客体只能被动地接受。如今，在大数据时代，这一影视制作、播出模式已被颠覆——观众想看什么，导演才拍什么。

在影视领域，大数据运用的成功案例当数美剧《纸牌屋》。该剧的制作方既不是电视台，也不是传统的电影公司，而是一家视频播放网站。2012 年，视频网站 Netflix 开始准备推出自制剧。在决定拍什么、怎么拍时，Netflix 抛开了传统的制作方式，启用大数据。通过在该网站上 3000 多万订阅用户每天的点击操作，如收藏、推荐、回放、暂停、搜索请求等，Netflix 进行精准分析，将这些数据用于倒推前台的影片生产。

通过对大数据的分析、挖掘，Netflix 发现，其用户中有很多人仍在点播 1990 年 BBC 经典老片《纸牌屋》。这些观众中，又有许多人喜欢导演大卫•芬奇，大多爱看演员凯文•史派西出演的电影。Netflix 大胆预测，一部影片如果同时满足这几个要素，就可能大卖。于是，《纸牌屋》出现了，并大获成功。整部剧集一次性在 Netflix 网站发布，供订阅者观看，完全颠覆了传统的每周一集的播出模式。

《纸牌屋》大获成功后，影视业兴起了利用大数据的浪潮。亚马逊等不少有实力的网站均开始通过大数据技术制作自制剧。



旅游：预先知道游客想去哪儿

前不久在法国举办的阿维尼翁论坛上，大数据也成为人们讨论的热词。

为吸引更多的外国游客，法国蓝色海岸区域旅游委员会和法国移动运营商 Orange 联合在蓝色海岸进行调研，对本区游客的游览路线、住宿、游客数量等进行测量和分析。试验期内，Orange 对超过 100 万个在漫游状态下的用户进行了分析。用户电话的定位系统提供了他们在该地区停留的时间，也较为准确地构建出了游客的活动范围。同时，Orange 公司与地区内的数据共享运营商 IGN 合作，有效收集到了该地区游客游览最多的地方、游客偏好的住所、到达该地区及离开的时间等。

这些微不足道的信息拼凑而成的大数据通过统计及图示的方式，为地区旅游发展提出了明确的修改方案，比如改善酒店位置、调整接待不同国籍外国游客的方式及地点、针对不同种类的旅游者制定相应类型的活动等。这些方案有效提高了地区的旅游服务。

同样的，法国旅游发展署也通过委托专业公司或到外国社交网络进行数据收集进行大数据分析。由于中国、日本游客近年来是法国旅游的主力军，法国旅游发展署在中国和日本的社交网络开展了相应的数据收集。

在线音乐：精确投放广告

在欧美国家，网络音乐下载需付费。在线收听免费，但需忍受音乐之间插播的广告。不合时宜的广告往往让用户大为扫兴。在大数据时代，这一状况将得到改善。一些在线音乐服务商通过收集用户的数据，如音乐类型偏好、收听音乐的场所、时间段等分析用户的口味，从而推送让用户感兴趣的广告，提高用户的体验度。

Pandora 电台是美国最流行的提供在线音乐服务的软件，拥有 7200 万名活跃用户。用户只要输入喜欢的歌曲或歌手，Pandora 电台就会为用户建立一个私人电台，不断播放风格相近的音乐。收听过程中，用户还可以选择“喜欢”“不喜欢”或者“我听腻了”，来对电台进行调整，使其播放的音乐更符合自己的口味。

基于大数据研究，今年 1 月，Pandora 电台推出“口味广告”，力求为用户插播“最适合”的广告。电台通过长期播放用户喜欢的音乐，进一步发掘每个用户喜欢的音乐类型，找到类似风格的广告进行投放。例如，在一个周末下午，用户正在收听激昂风格的音乐，电台会考虑投放一个关于波多黎各冒险游的广告；而对于一个周一早晨在办公室听类似激昂风格音乐的用户来说，电台也许会投放一个传统巴黎之旅的广告。

艺术品市场：预测交易数量和市场变化

2013 年 3 月，德勤和 ArtTactic 联合发布了 2013 艺术品网上交易报告。该报告显示，艺术品在线拍卖将成为新增长点。过去的几年里，超过 300 家在线艺术品风投公司成立，至少 71%的艺术收藏家在线购买过艺术品。随着这一数字的增长，新的数据将会不断产生。大数据对于其他行业的引领和预测，在艺术市场方面一样适用。

艺术品行业的大数据主要包括用户、内容和渠道三方面。艺术品数据公司 Artnet 建立了覆盖美术、设计和装饰艺术的丰富数据库。签约用户可以通过搜集的艺术品交易记录，分析艺术品市场变化。比如将一些艺术家的信息选出，综合出一个流派的指数，从中观察交易数量的变化。这为艺术品研究报告的撰写提供了准确的数据。据相关报道，以过去 10 年来的艺术品交易市场情况为对象，通过大数据挖掘，发现占市场交易最大份额的是现代派和印象派画作，其中，现代派占比 34%，印象派则为 24%，二者抢占了全球艺术品市场的大半江山。

时装设计：社交媒体抢走前排座椅

大数据对时装设计也产生了变革性的影响。在时装行业，失策的代价将会非常高昂，而时装成功与否，往往取决于是否选择了合适的图案、颜色、面料、形状、尺寸等，而这些都属于大数据的范畴。

时装行业的大数据源自无孔不入的社交媒体。每天，全球有超过 10 亿人活跃在社交网络上。每天都有上百万人在社交媒体上评论、分享、发微博，讨论什么是潮流。

在大数据时代，最热门的时尚潮流不再专属于 T 型台。越来越多的知名设计师、品牌和零售商开始利用社交网络让公众参与到设计当中，而此前，这一行业只对时尚精英开放。例如，梅赛德斯-奔驰国际时装周已不再为时尚精英所独享。在过去的几个季节，越来越多的顶级设计师和品牌都在时装秀之前和期间在网上发布全新的设计，如奥斯卡•德拉伦塔在 Instagram 上发布了最新的高级女装成衣系列，巴宝莉的官方微博账号在模特们走秀之前就发布了后台的照片……顶级买家和时尚杂志主编也感受到社交媒体的影响。原本属于他们的前排座椅已被流行博客写手、拥有大批微博粉丝的摄影师和网络红人所占据，他们对大众的时尚影响力远大于传统的精英人群。

电子游戏：让玩家参与创作

如今，无论是社交游戏、手机游戏还是主机游戏，数据和分析在开发过程中都发挥着重要作用。通过大数据分析每日、每月活跃用户数，玩家支付费用和游戏时间等，游戏设计者不仅能使已存在游戏的客户体验度得到提升，也能为新游戏的推出减免不必要的风险。

首次将数据分析的概念引进游戏领域的是社交游戏。Zynga、Playfish 等早期采纳者由于针对客户体验数据采取对策而获得巨大收获，并通过增值性的调整留存了更多玩家，掌握了吸引新用户的方法。Zynga 公司在发行游戏《城市小镇》之初，通过数据分析，发现新玩家很难完成游戏的初级任务，及时调整了关卡设计，在不丢失游戏魅力的基础上使任务简单化，并根据用户数据完善游戏设计，使游戏成功吸引更多玩家。

就此，游戏设计者斯科特•休梅克表示，在游戏发行之前，游戏设计者很难对游戏图形、级别设计及趣味性、吸引力等进行准确评估，所以游戏设计的推行、测试和调整是重要的一环。这个过程离不开玩家的参与，更离不开大数据。

【原文链接】 【回到目录】



【玩转·大数据】

David Brooks：“大数据”时代，什么是数据分析做不了的？


Daivd Brooks：《纽约时报》文化记者

“

目前这一历史时期最大的创新就在于，我们的生活现在由收集数据的计算机调控着。在这个时代，头脑无法理解的复杂情况，数据可以帮我们解读其中的含义。数据可以弥补我们对直觉的过分自信，数据可以减轻欲望对知觉的扭曲程度。

”

不久之前我曾与一位大型银行的首席执行官一同用餐。他正在考虑是否要退出意大利市场，因为经济形势不景气，而且未来很可能出现一场欧元危机。

这位 CEO 手下的经济学家描绘出一片惨淡的景象，并且计算出经济低迷对公司意味着什么。但是最终，他还是在自己价值观念的指引下做出了决定。

这家银行在意大利已经有了几十年的历史。他不希望意大利人觉得他的银行只能同甘不能共苦。他不希望银行的员工认为他们在时局艰难之际会弃甲而逃。他决定留在意大利，不管未来有什么危机都要坚持下去，即便付出短期代价也在所不惜。

做决策之时他并没有忘记那些数据，但最终他采用了另一种不同的思维方式。当然，他是正确的。商业建立在信任之上。信任是一种披着情感外衣的互惠主义。在困境中做出正确决策的人和机构能够赢得自尊和他人的尊敬，这种感情上的东西是非常宝贵的，即便它不能为数据所捕捉和反映。

这个故事反映出了数据分析的长处和局限。目前这一历史时期最大的创新就在于，我们的生活现在由收集数据的计算机调控着。在这个时代，头脑无法理解的复杂情况，数据可以帮我们解读其中的含义。数据可以弥补我们对直觉的过分自信，数据可以减轻欲望对知觉的扭曲程度。

但有些事情是“大数据”不擅长的，下面我会一一道来：

数据不懂社交。大脑在数学方面很差劲（不信请迅速心算一下 437 的平方根是多少），但是大脑懂得社会认知。人们擅长反射彼此的情绪状态，擅长侦测出不合作的行为，擅长用情绪为事物赋予价值。

计算机数据分析擅长的是测量社会交往的“量”而非“质”。网络科学家可以测量出你在76%的时间里与 6 名同事的社交互动情况，但是他们不可能捕捉到你心底对于那些一年才见两次的儿时玩伴的感情，更不必说但丁对于仅有两面之缘的贝阿特丽斯的感情了。因此，在社交关系的决策中，不要愚蠢到放弃头脑中那台充满魔力的机器，而去相信你办公桌上的那台机器。

数据不懂背景。人类的决策不是离散的事件，而是镶嵌在时间序列和背景之中的。经过数百万年的演化，人脑已经变得善于处理这样的现实。人们擅长讲述交织了多重原因和多重背景的故事。数据分析则不懂得如何叙事，也不懂得思维的浮现过程。即便是一部普普通通的小说，数据分析也无法解释其中的思路。

数据会制造出更大的“干草垛”。这一观点是由纳西姆•塔勒布（Nassim Taleb，著名商业思想家，著有《黑天鹅：如何应对不可知的未来》等书作）提出的。随着我们掌握的数据越来越多，可以发现的统计上显著的相关关系也就越来越多。这些相关关系中，有很多都是没有实际意义的，在真正解决问题时很可能将人引入歧途。这种欺骗性会随着数据的增多而指数级地增长。在这个庞大的“干草垛”里，我们要找的那根针被越埋越深。大数据时代的特征之一就是，“重大”发现的数量被数据扩张带来的噪音所淹没。

大数据无法解决大问题。如果你只想分析哪些邮件可以带来最多的竞选资金赞助，你可以做一个随机控制实验。但假设目标是刺激衰退期的经济形势，你就不可能找到一个平行世界中的社会来当对照组。最佳的经济刺激手段到底是什么？人们对此争论不休，尽管数据像海浪一般涌来，就我所知，这场辩论中尚未有哪位主要“辩手”因为参考了数据分析而改变立场的。



数据偏爱潮流，忽视杰作。当大量个体对某种文化产品迅速产生兴趣时，数据分析可以敏锐地侦测到这种趋势。但是，一些重要的（也是有收益的）产品在一开始就被数据摈弃了，仅仅因为它们的特异之处不为人所熟知。

数据掩盖了价值观念。我最近读到一本有着精彩标题的学术专著——《“原始数据”只是一种修辞》。书中的要点之一就是，数据从来都不可能是“原始”的，数据总是依照某人的倾向和价值观念而被构建出来的。数据分析的结果看似客观公正，但其实价值选择贯穿了从构建到解读的全过程。

这篇文章并不是要批评大数据不是一种伟大的工具。只是，和任何一种工具一样，大数据有拿手强项，也有不擅长的领域。正如耶鲁大学的爱德华•图弗特教授（Edward Tufte）所说：“这个世界的有趣之处，远胜任何一门学科。”

（文章来源：果壳网）

【原文链接】　【回到目录】



Tim Harford：大数据，还是大错误？


Tim Harford：英国经济学家，记者

“

大数据就好像是蛮荒的美国西部。那些头脑灵活野心勃勃的人会想尽办法利用一切可能的工具，从这些数据中淘出点值钱的东西来，这很酷。但目前我们做的还有些盲目。

”

大数据是对于大规模现象的一种模糊的表达。这一术语如今已经被企业家、科学家、政府和媒体炒得过热。

五年前，谷歌的一个研究小组在全球顶级的科学杂志《自然》上宣布了一个令人瞩目的成果。该小组可以追踪美国境内流感的传播趋势，而这一结果不依赖于任何医疗检查。他们的追踪速度甚至比疾控中心(CDC)要快的多。谷歌的追踪结果只有一天的延时，而 CDC 则需要汇总大量医师的诊断结果才能得到一张传播趋势图，延时超过一周。谷歌能算的这么快，是因为他们发现当人们出现流感症状的时候，往往会跑到网络上搜索一些相关的内容。

“谷歌流感趋势”不仅快捷、准确、成本低廉，而且没有使用什么理论。谷歌的工程师们不用费劲的去假设哪些搜索关键字（比如“流感症状”或者“我身边的药店”）跟感冒传染有相关性。他们只需要拿出来自己网站上 5000 万个最热门的搜索字，然后让算法来做选择就行了。

谷歌流感趋势的成功，很快就成为了商业、技术和科学领域中最新趋势的象征。兴奋的媒体记者们不停的在问，谷歌给我们带来了什么新的科技？

在这诸多流行语中，“大数据”是一个含糊的词汇，常常出现于各种营销人员的口中。一些人用这个词来强调现有数据量的惊人规模——大型粒子对撞机每年会产生 15PB 的数据，相当于你最喜欢的一首歌曲重复演奏 15000 年的文件大小。



然而在“大数据”里，大多数公司感兴趣的是所谓的“现实数据”，诸如网页搜索记录、信用卡消费记录和移动电话与附近基站的通信记录等等。谷歌流感趋势就是基于这样的现实数据，这也就是本文所讨论的一类数据。这类数据集甚至比对撞机的数据规模还要大（例如 Facebook），更重要的是虽然这类数据的规模很大，但却相对容易采集。它们往往是由于不同的用途被搜集起来并杂乱的堆积在一起，而且可以实时的更新。我们的通信、娱乐以及商务活动都已经转移到互联网上，互联网也已经进入我们的手机、汽车甚至是眼镜。因此我们的整个生活都可以被记录和数字化，这些在十年前都是无法想象的。

大数据的鼓吹者们提出了四个令人兴奋的论断，每一个都能从谷歌流感趋势的成功中印证：

1）数据分析可以生成惊人准确的结果；

2）因为每一个数据点都可以被捕捉到， 所以可以彻底淘汰过去那种抽样统计的方法；

3）不用再寻找现象背后的原因，我们只需要知道两者之间有统计相关性就行了；

4）不再需要科学的或者统计的模型，“理论被终结了”。《连线》杂志 2008 年的一篇文章里豪情万丈的写到：“数据已经大到可以自己说出结论了”。

不幸的是，说的好听一些，上述信条都是极端乐观和过于简化了。如果说的难听一点，就像剑桥大学公共风险认知课的 Winton 教授（类似于国内的长江学者——译者注）David Spiegelhalter 评论的那样，这四条都是“彻头彻尾的胡说八道”。

在谷歌、Facebook 和亚马逊这些公司不断通过我们所产生的数据来理解我们生活的过程中，现实数据支撑起了新互联网经济。爱德华•斯诺登揭露了美国政府数据监听的规模和范围，很显然安全部门同样痴迷从我们的日常数据中挖掘点什么东西出来。

咨询师敦促数据小白们赶紧理解大数据的潜力。麦肯锡全球机构在一份最近的报告中做了一个计算，从临床试验到医疗保险报销到智能跑鞋，如果能把所有的这些健康相关的数据加以更好的整合分析，那么美国的医疗保险系统每年可以节省 3000 亿美金的开支，平均每一个美国人可以省下 1000 美元。

虽然大数据在科学家、企业家和政府眼里看起来充满希望，但如果忽略了一些我们以前所熟知的统计学中的教训，大数据可能注定会让我们失望。

Spiegelhalter 教授曾说到：“大数据中有大量的小数据问题。这些问题不会随着数据量的增大而消失，它们只会更加突出。”



在那篇关于谷歌流感趋势预测的文章发表 4 年以后，新的一期《自然杂志消息》报道了一则坏消息：在最近的一次流感爆发中谷歌流感趋势不起作用了。这个工具曾经可靠的运作了十几个冬天，在海量数据分析和不需要理论模型的条件下提供了快速和准确的流感爆发趋势。然而这一次它迷路了，谷歌的模型显示这一次的流感爆发非常严重，然而疾控中心在慢慢汇总各地数据以后，发现谷歌的预测结果比实际情况要夸大了几乎一倍。

问题的根源在于谷歌不知道（一开始也没法知道）搜索关键词和流感传播之间到底有什么关联。谷歌的工程师们没有试图去搞清楚关联背后的原因。他们只是在数据中找到了一些统计特征。他们更关注相关性本身而不是相关的原因。这种做法在大数据分析中很常见。要找出到底是什么原因导致了某种结果是很困难的，或许根本不可能。而发现两件事物之间的相关性则要简单和快速的多。就像 Viktor Mayer-Schönberger 和 Kenneth Cukier 在《大数据》这本书中形容的那样：“因果关系不能被忽略，然而曾作为所有结论出发点的它已经被请下宝座了。”

这种不需要任何理论的纯粹的相关性分析方法，其结果难免是脆弱的。如果你不知道相关性背后的原因，你就无法得知这种相关性在什么情况下会消失。谷歌的流感趋势出错的一种解释是，2012 年 12 月份的媒体上充斥着各种关于流感的骇人故事，看到这些报道之后，即使是健康的人也会跑到互联网上搜索相关的词汇。还有另外一种解释，就是谷歌自己的搜索算法，在人们输入病症的时候会自动推荐一些诊断结果进而影响到了用户的搜索和浏览行为。这就好像在足球比赛里挪动了门柱一样，球飞进了错误的大门。

谷歌将使用新的数据再次校准流感趋势这个产品，重新来过。这当然是正确的做法。能够有更多的机会让我们简捷的采集和处理大规模的数据，这当然有一百个理由让人兴奋。然而我们必须从上述例子中汲取足够的教训，才能避免重蹈覆辙。

统计学家们过去花了 200 多年，总结出了在认知数据的过程中存在的种种陷阱。如今数据的规模更大了，更新更快了，采集的成本也更低了。但我们不能掩耳盗铃，假装这些陷阱都已经被填平了，事实上它们还在那里。

在 1936 年，民主党人 Alfred Landon（艾尔弗雷德·兰登）与当时的总统 Franklin Delano Roosevelt（富兰克林·罗斯福——译者注）竞选下届总统。《读者文摘》这家颇有声望的杂志承担了选情预测的任务。当时采用的是邮寄问卷调查表的办法，调查人员雄心勃勃，计划寄出 1000 万份调查问卷，覆盖四分之一的选民。可以预见，洪水般寄回的邮件将超乎想象，然而《文摘》似乎还乐在其中。8 月下旬的时候他们写到：“从下周起，



1000 万张问卷的第一批回执将会到达，这将是后续邮件洪峰的开始。所有这些表格都会被检查三次，核对，交叉存档五份，然后汇总。”

最终《文摘》在两个多月里收到了惊人的 240 万份回执，在统计计算完成以后，杂志社宣布 Landon 将会以 55 比 41 的优势击败 Roosevelt 赢得大选，另外 4%的选民则会投给第三候选人。

然而真实选举结果与之大相径庭：Roosevelt 以 61 比 37 的压倒性优势获胜。让《读者文摘》更没面子的是，观点调查的先创人 George Gallup 通过一场规模小的多的问卷，得出了准确得多的预测结果。Gallup 预计 Roosevelt 将稳操胜券。显然，Gallup 先生有他独到的办法。而从数据的角度来看，规模并不能决定一切。

观点调查是基于对投票人的的大范围采样。这意味着调查者需要处理两个难题：样本误差和样本偏差。

样本误差是指一组随机选择的样本观点可能无法真实的反映全部人群的看法。而误差的幅度，则会随着样本数量的增加而减小。对于大部分的调查来说，1000 次的访谈已经是足够大的样本了。而据报道 Gallup 先生总共进行了 3000 次的访谈。

就算 3000 次的访谈已经很好了，那 240 万次不是会更好吗？答案是否定的。样本误差有个更为危险的朋友：样本偏差。样本误差是指一个随机选择的样本可能无法代表所有其他的人；而样本偏差则意味着这个样本可能根本就不是随机选择的。George Gallup 费了很大气力去寻找一个没有偏差的样本集合，因为他知道这远比增加样本数量要重要的多。

而《读者文摘》为了寻求一个更大的数据集，结果中了偏差样本的圈套。他们从车辆注册信息和电话号码簿里选择需要邮寄问卷的对象。在 1936 年那个时代，这个样本群体是偏富裕阶层的。而且 Landon 的支持者似乎更乐于寄回问卷结果，这使得错误更进了一步。这两种偏差的结合，决定了《文摘》调查的失败。Gallup 每访谈一个人，《文摘》对应的就能收到 800 份回执。如此大规模而精确的调查最终却得出一个错误的结果，这的确让人难堪不已。

如今对大数据的狂热似乎又让人想起了《读者文摘》的故事。现实数据的集合是如此混乱，很难找出来这里面是否存在样本偏差。而且由于数据量这么大，一些分析者们似乎认定采样相关的问题已经不需要考虑了。而事实上，问题依然存在。



《大数据》这本书的联合作者，牛津大学互联网中心的 Viktor Mayer-Schönberger 教授，曾告诉我，他最喜欢的对于大数据集合的定义是“N=所有”，在这里不再需要采样，因为我们有整个人群的数据。就好比选举监察人不会找几张有代表性的选票来估计选举的结果，他们会记点每一张选票。当“N=所有”的时候确实不再有采样偏差的问题，因为采样已经包含了所有人。

但“N=所有”这个公式对大多数我们所使用的现实数据集合都是成立的吗？恐怕不是。“我不相信有人可以获得所有的数据”，伦敦大学学院的计算机学家和统计学教授 Patrick Wolfe 说。

推特(Twitter)就是一个例子。理论上说你可以存储和分析推特上的每一条记录，然后用来推导出公共情绪方面的一些结论（实际上，大多数的研究者使用的都是推特提供的一个名为“消防水龙带”的数据子集）。然而即使我们可以读取所有的推特记录，推特的用户本身也并不能代表世界上的所有人。（根据 Pew 互联网研究项目的结果，在 2013 年，美国的推特中年轻的，居住在大城市或者城镇的，黑色皮肤的用户比例偏高）

我们必须要搞清楚数据中漏掉了哪些人和哪些事，尤其当我们面对的是一堆混乱的现实数据的时候。Kaiser Fung 是一名数据分析师和《数字感知》这本书的作者，他提醒人们不要简单的假定自己掌握了所有有关的数据：“N=所有，常常是对数据的一种假设，而不是现实”。

在波士顿有一款智能手机应用叫做“颠簸的街道”，这个应用利用手机里的加速度感应器来检查出街道上的坑洼，而有了这个应用市政工人就可以不用再去巡查道路了。波士顿的市民们下载这个应用以后，只要在城市里开着车，他们的手机就会自动上传车辆的颠簸信息并通知市政厅哪里的路面需要检修了。几年前还看起来不可思议的事情，就这样通过技术的发展，以信息穷举的方式得以漂亮的解决。波士顿市政府因此骄傲的宣布，“大数据为这座城市提供了实时的信息，帮助我们解决问题并做出长期的投资计划”。

“颠簸的街道”在安装它的设备中所产生的，是一个关于路面坑洼的地图。然而从产品设计一开始这张地图就更偏向于年轻化和富裕的街区，因为那里有更多的人使用智能手机。“颠簸的街道”的理念是提供关于坑洼地点的“N=所有”的信息，但这个“所有”指的是所有手机所能记录的数据，而不是所有坑洼地点的数据。就像微软的研究者 Kate Crawford 指出的那样，现实数据含有系统偏差，人们需要很仔细的考量才可能找到和纠正



这些偏差。大数据集合看起来包罗万象，但“N=所有”往往只是一个颇有诱惑力的假象而已。

当然这个世界的现实是如果你能靠某个概念挣到钱，就没人会关心什么因果关系和样本偏差。全世界的公司在听到美国折扣连锁店 Target 的传奇式成功（由纽约时报的 Charles Duhigg 在 2012 年报道出来）以后估计都要垂涎三尺。Duhigg 解释了 Target 公司是如何从它的顾客身上搜集到大量的数据并熟练的加以分析。它对顾客的理解简直是出神入化。

Duhigg 讲的最多的故事是这样的：一名男子怒气冲冲的来到一家明尼苏达附近的 Target 连锁店，向店长投诉该公司最近给他十几岁的女儿邮寄婴儿服装和孕妇服装的优惠券。店长大方的向他道了歉。可不久后店长又收到这名男子的电话要求再次道歉——只是这一次对方告知那个少女确实怀孕了。在她的父亲还没有意识到的时候，Target 通过分析她购买无味湿纸巾和补镁药品的记录就猜到了。

这是统计学的魔法吗？或许还有更世俗一点的解释。

Kaiser Fung 在帮助零售商和广告商开发类似的工具上有着多年的经验，他认为“这里面存在一个严重的虚假正面效应的问题”。他指的是我们通常都没能够听到的无数的反面故事，在那些例子里没有怀孕的妇女们也收到了关于婴儿用品的优惠券。

如果只听 Duhigg 讲的故事，你可能很容易就觉得 Target 的算法是绝对可靠的——每个收到婴儿连体服和湿纸巾购物券的人都是孕妇。这几乎不可能出错。但实际上孕妇能收到这些购物券可能仅仅是因为 Target 给所有人都寄了这种购物券。在相信 Target 那些读心术般的故事之前，你应当问问他们的命中率到底有多高。

在 Charles Duhiggs 的描述中，Target 公司会在给你的购物券中随机性的掺杂一些无关的东西，比如酒杯的券。否则的话孕妇们可能会发现这家公司的计算机系统在如此深入的探测她们的隐私，进而感到不安。

Fung 对此则有另外的解释，他认为 Target 这样做并不是因为给孕妇寄一份满是婴儿用品的购物手册会让人起疑，而是由于这家公司本来就知道这些手册会被寄给很多根本没有怀孕的妇女。

以上这些观点并不意味着数据分析一无是处，相反它可能是有高度商业价值的。即使能够把邮寄的准确度提高那么一点点，都将是有利可图的。但能赚钱并不意味着这种工具无所不能、永远正确。



一位名叫 John Ioannidis 的传染病学家在 2005 年发表了一篇论文，题目叫“为什么大多数被发表的研究结果都是错误的”，标题言简意赅。他的论文中一个核心的思想就是统计学家们所称的“多重比较问题”。

当我们审视数据当中的某个表象的时候，我们常常需要考虑这种表象是否是偶然产生的。如果这种表象看起来不太可能是随机产生的时候，我们就称它是“统计上显著的”。

当研究者面对许多可能的表象时，多重比较错误就可能发生。假设有一个临床试验，我们让部分小学生服用维他命而给其他小学生安慰剂。怎么判断这种维他命的效果？这完全取决于我们对“效果”的定义。研究者们可能会考察这些儿童的身高、体重、蛀牙的概率、课堂表现、考试成绩甚至是 25 岁以后的收入或者服刑记录（长期追踪观察）。然后是综合比较：这种维他命是对穷困家庭的孩子有效，还是对富裕家庭的有效？对男孩有效，还是女孩？如果做足够多的不同的相关性测试，偶然产生的结果就会淹没真实的发现。

有很多办法可以解决上述的问题，然而在大数据中这种问题会更加严重。因为比起一个小规模的数据集合来说，大数据的情况下有太多可以用作比较的标准。如果不做仔细的分析，那么真实的表象与虚假表象之比——相当于信号噪声比——很快就会趋近于 0.

更糟的是，我们之前会用增加过程透明度的办法来解决多重比较的问题，也就是让其他的研究者也知道有哪些假设被测试过了，有哪些反面的试验结果没有被发表出来。然而现实数据几乎都不是透明的。亚马逊和谷歌，Facebook 和推特，Target 和 Tesco，这些公司都没打算过跟你我分享他们的所有数据。

毫无疑问，更新、更大、更廉价的数据集合以及强大的分析工具终将产生价值。也确实已经出现了一些大数据分析的成功实例。剑桥的 David Spiegelhalter 提到了谷歌翻译，这款产品统计分析了人类已经翻译过的无数文档，并在其中寻找出可以自己复制的模式。谷歌翻译是计算机学家们所谓的“机器学习”的一个应用，机器学习可以在没有预先设定编程逻辑的条件下计算出惊人的结果。谷歌翻译是目前所知的最为接近“无需理论模型、纯数据驱动的算法黑盒子”这一目标的产品。用 Spiegelhalter 的话来说，它是“一个令人惊讶的成就”。这一成就来自于对海量数据的聪明的处理。

然而大数据并没有解决统计学家和科学家们数百年来所致力的一些难题：对因果关系的理解，对未来的推演，以及如何对一个系统进行干预和优化。

伦敦皇家学院的 David Hand 教授讲过一句话，“现在我们有了一些新的数据来源，但是没有人想要数据，人们要的是答案”。



要使用大数据来得到这样的答案，还需要在统计学的方法上取得大量长足的进展。

伦敦大学学院的 Patrick Wolfe 说，“大数据就好像是蛮荒的美国西部。那些头脑灵活、野心勃勃的人会想尽办法利用一切可能的工具，从这些数据中淘出点值钱的东西来，这很酷。但目前我们做的还有些盲目。”

统计学家们正争先恐后的为大数据开发新的工具。这些新的工具当然很重要，但它们只有在吸取而不是遗忘过去统计学精髓的基础上才能成功。

最后，我们再回头来看看大数据的四个基础信条。其一，如果简单的忽略掉那些反面的数据，比如 Target 的怀孕预测算法，那么我们很容易就会过高的估计算法的精确度。其二，如果我们在一个固定不变的环境里做预测，你可以认为因果关系不再重要。而当我们处在一个变化的世界中（例如流感趋势预测所遇到的那样），或者是我们自己就想要改变这个环境，这种想法就很危险了。其三，“N=所有”，以及采样偏差无关紧要，这些前提在绝大多数的实际情况下都是不成立的。最后，当数据里的假像远远超过真相的时候，还持有“数据足够大的时候，就可以自己说出结论了”这种观点就显得过于天真了。

大数据已经到来，但它并没有带来新的真理。现在的挑战是要吸取统计学中老的教训，在比以前大得多的数据规模下去解决新的问题、获取新的答案。

【原文链接】　　【回到目录】



经济学人：数据，无处不在的数据

The Economist

Clair Han：译言翻译

“

从信息匮乏到信息过剩，这一转变带来了广泛的影响。微软研究及策略主管克瑞格·蒙迪（Craig Mundie）说‘我们正在看到的是根据数据而形成经济的能力，对于我来说，这是社会层面甚至是宏观经济层面的一次大改变’。数据正逐渐成为商业新的原始资料：基本与资本和劳动力相同的经济投入。‘每天我醒来都会问自己：“怎样才能更好地使数据流动，更好地处理数据，更好地分析数据呢？”’沃尔玛首席信息官洛林·福特（Rollin Ford）说。

”

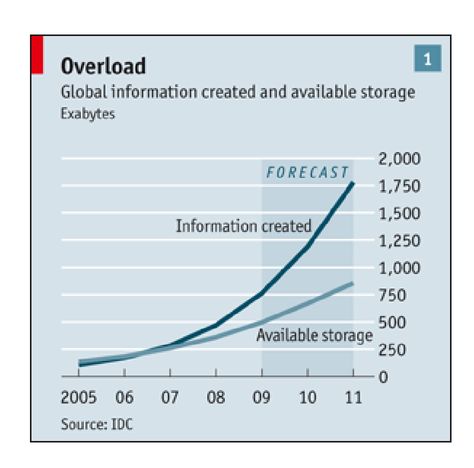
自 2000 年斯隆数字巡天开始工作，它位于墨西哥的望远镜前几周收集到的数据就远远超过了天文史上所获得的数据总和。十年后的今天，它收集到的信息已达 140 万亿字节。其继任者，应于 2016 年在智利投入使用的大型综合巡天望远镜，每五天就能收集到 140 万亿字节的数据。

地球上也有如此巨大的信息量。零售巨头沃尔玛每小时都要处理 100 多万笔交易，为数据库大概上传 2500 兆数据，相当于美国国会图书馆存书数的 167 倍（数据如何量化请见文章 article）。社交网站脸谱网，储存了 400 亿张照片。而解析人类基因则要分析 30 亿碱基对——十年前第一次进行这项工作时花费了十年时间，而 2003 时，只用了一周时间就完成了。

所有的例子都说明了一件事：这个世界上数据量之多难以想象，而它还在快速增长着。数据多了，我们可以做很多之前不能做的事情，比如发现经济趋势、防治疾病、打击犯罪等等。若管理得当，数据能发现经济价值新来源、为科学提供新思路、监督政府履行职责。



但是，数据也产生了新麻烦。除了捕捉、处理以及分享数据的工具——例如传感器、电脑、手机等等，数据量已远远超过可用储存空间所能承受的范围（见表一）。此外，随着信息成倍的增长，并被更广泛的传播，保证数据安全、保护隐私已经越来越难了。



美国约翰霍金斯大学天体物理学家亚历克斯•萨雷（Alex Szalay）指出数据大量增长使得获取数据越来越难。“该如何弄懂这些数据呢？人们要担心的不是培养科学家，而是整个下一代，包括从政从商的人们。”他说。

“大量信息的存在使得我们所处的时代与以往不同。”美国国际商用机器公司（IBM）的詹姆士•科塔达（James Cortada）说，詹姆士写过十好几本关于社会信息史的书。来自伯克利加利福尼亚大学的计算机科学家乔•海勒斯坦（Joe Hellerstein）称我们这个时代为“数据的工业革命”。从商业到科学，从政府到艺术，数据无时不刻不在发挥作用。科学家和计算机工程师为这一现象创造了一个新术语：“大数据”。

从认知论角度说，信息是数据的集合，知识是信息的集合。但在这份特殊的报告中，“数据”和“信息”可以相互替换，因为这两者已经越来越难以区分了。若有足够的原始数据，如今强大的计算机和算法可以提供之前未被发现的新思路。

帮助各个组织理解各自激增的数据即信息管理业务，这一业务如今正飞速发展。美国国际商用机器公司、微软公司以及介于两者之间的思爱普（SAP）公司已花费 150 亿美元购买数据处理及分析公司。据估计，这一行业的价值已超过 1 千亿美元并以将近每年 10%的速度增长，基本相当于整个软件行业整体增速的两倍。

首席信息官们（CIOs）在行政系统中的地位更加重要了，此外，数据科学家这一新兴职业衍生出来了，数据科学家集软件程序员、统计员以及演讲者于一身，在海量的数据里找出隐藏着的金砖。哈尔•范里安（Hal Varian）是谷歌的首席经济专家，他预言统计员的工作



将成为“最性感”的工作。在他看来，数据可被广泛应用，缺少的只是从中汲取智慧的能力。

什么都更多了

信息爆炸的原因有很多，首当其冲的是科技。数字装置的各种能力不断提升，价格却不断下降，传感器还有小工具将许多之前无法接触到的数据数字化。而且，许多人都有更加强大的工具。例如，全球有 46 亿手机用户（由于有些人拥有不止一部手机，所以 68 亿手机用户这一数据并不精准）和 10—20 亿互联网用户。

此外，现在还有许多人与信息进行互动。科塔达先生指出，1990 年到 2005 年间，全球超过 10 亿人成为中产阶级，这些人变得富有的同时，也更有文化，这也促进了信息量的增长。这一点不仅体现在政治、经济上，也体现在法律上。“度量革命是科技革命的先导，”纽约大学经济学博士希南•阿拉尔（Sinan Aral）说。如同显微镜发现了细菌从而革新了生物学、电子显微镜改变了物理学，社会科学因这些数据正发生着翻天覆地的变化，他解释道。从前，研究者从个体层面理解人类行为，现在，他们可以从群体层面进行研究。

每五年，数字信息翻十倍。计算机行业公认的金科玉律摩尔定律认为，价格不变的情况下，每十八个月计算机芯片的处理能力以及储存能力差不多会翻一倍。微软程序也会变得更好。普林斯顿大学计算机科学家爱德华•费尔顿（Edward Felten）认为，算法的进步促使计算机应用程序几十年来和摩尔定律同等重要。

大量的信息被共享。通信设备制造商思科（Cisco）称，到 2013 年，互联网上的数据流通量将达到 667 艾字节（百亿亿字节）。而且，数据会不断增长，增速比互联网承载能力还要快。

人们常抱怨自己被湮没在信息中。1917 年，康涅狄格州制造公司经理曾抱怨电话的影响：“浪费时间浪费钱还造成混乱。”然而，现在发生的一切早已不再是单纯的量的增长。量变已开始引起质变。

从信息匮乏到信息过剩，这一转变带来了广泛的影响。微软研究及策略主管克瑞格•蒙迪（Craig Mundie）说“我们正在看到的是根据数据而形成经济的能力，对于我来说，这是社会层面甚至是宏观经济层面的一次大改变”。数据正逐渐成为商业新的原始资料：基本与资本和劳动力相同的经济投入。“每天我醒来都会问自己：‘怎样才能更好地使数据流

动，更好地处理数据，更好地分析数据呢？’”沃尔玛首席信息官洛林•福特（Rollin Ford）说。

成熟的定量分析被用于生活中的很多方面，并不像之前只被用于设计导弹弹道或者金融对冲策略中。预测（Farecast）是微软搜索引擎必应的一部分，它可以在分析过 2250 亿个航班及机票记录后，为使用者提供是购买还是等价格降低的建议。同一理念也被扩展应用于酒店房间、汽车以及相似的预定中。个人理财网站以及银行集合了客户数据用于揭示宏观经济走势，这使他们能在各自的领域中提供一些配套业务。喜欢捣弄数字的人甚至发现了日本相扑比赛中的假赛。

提炼成金

“数据排放”，即互联网用户留下的有价值的点击记录，正逐渐成为互联网经济的中流砥柱。谷歌的索引擎就是一个例子，它一定程度上是通过某个条目的点击数来帮助人们确定搜索内容的相关度。如果第八个搜索条目有最多的人惦记，那么算法就会把它往前提。

随着世界变得越来越数字化，整合、分析数据也给其他领域带来了巨大的好处。例如，微软的蒙迪先生和谷歌的老板埃里克•施密特（Eric Schmidt）受总统之邀进行美国医疗保健系统改革。“任务开始初期，埃里克和我对对方说：‘看吧，如果你想真正的改革医疗保健系统，基本上就得围绕数据建造一种医疗保健经济，而这些数据得和人有关’，”蒙迪先生说。“你不应把数据当做提供医疗服务的“废物”，而要当做弄清楚如何改进医疗保健各方面的重要资产。这有点逆向思维。”

可以确定的是，数字记录可以让医生的工作更轻松、降低医疗成本、提高医疗质量。此外，数据整合还可以发现一些药物间的有害作用、确认最有效的治疗方案并在病症出现之前预测疾病。计算机已经在尝试着做这些事情，但是它们需要更加明确的指令。在大数据时代，几乎所有的联系都要自行出现。有时，数据显示出的东西会超过所需。例如，加利福尼亚州的奥克兰市在一个死人网站 Oakland Crimespotting 上发布信息，公布了逮捕的时间地点。一段时间的点击显示，警察每天都在一条繁忙的街道上抓捕卖淫人员，只有周三不去，也许这正是警察的秘密。

但大数据还带来了更多更严重的后果。比如说，众所周知最近的经济危机中，尽管各家银行和评级机构能够获取大量的信息，但是仍未能预测现实世界中的经济危机。这是由大数据引起的第一场危机，之后还会有更多。

处理信息的方式会在方方面面影响生活。20 世纪之交，随电报和电话而来的新的信息流使得大规模生产成为可能。现在，能够接触丰富的数据使得公司可以为世界各地的小型利基市场服务。以前，经济生产依靠工厂，管理者们钻研每一台机器，研究每一个生产步骤以提高工作效率。现在，统计员们从商业信息中挖掘新思路。

“以数据为导向的经济还处于初始阶段，”微软的蒙迪先生说。“人们可以看到它大概的轮廓，但现在还无法完全在技术层面，基础设施层面甚至商业模式意义层面理解它。”本专题报道指向它浮现的开端。

【原文链接】　　【回到目录】



【玩坏·大数据】

Kate Crawford：大数据真有这么神奇吗？


Kate Crawford：微软研究院首席研究员、麻省理工学院公民媒体中心客座教授

“大数据的鼓吹者希望人们相信，在一行行的代码和庞大数据库的背后存在着有关人类行为模式的客观、普遍的洞察，不管是消费者的支出规律、犯罪或恐怖主义行动、健康习惯，还是雇员的生产效率。但是许多大数据的传道者不愿正视其不足。数字无法自己说话，而数据集——不管它们具有什么样的规模——仍然是人类设计的产物。”

“大数据”是当前的时髦术语，是技术界用来解决世界上最难处理的问题的全能办法。这个术语一般用来描述对海量信息进行分析，从而发现规律、收集感悟和预言复杂问题答案的艺术与科学。它也许听起来有些乏味，但是从制止恐怖分子到消除贫穷，再到拯救地球，对于大数据的鼓吹者来说，没有什么问题是解决不了的。

维克托•梅耶-舍恩伯格和肯尼思•丘基尔在有着朴素书名的《大数据：一次将改变我们生活、工作和思考方式的革命》一书中欢呼道：“对社会的好处将是无穷无尽的，因为大数据在一定程度上将解决迫在眉睫的全球问题，如处理气候变化、根除疾病以及促进善政和经济发展等。”

只要有足够多的数据可以处理——不管是你的 iPhone 上的数据、杂货店购物状况、在线约会网站个人简介或者是整个国家的匿名健康记录，利用对这些原始数据进行解码的计算能力，人们可以获得数不胜数的洞察。甚至连奥巴马政府也已经赶上了这股潮流，并在 5 月



9 日向企业家、研究人员和公众“破天荒”地发布了大量“以前难以获取或难以管理的数据”。

然而，大数据真的完全像人们吹嘘的那样吗？人们能相信如此众多的 1 和 0 将能揭示人类行为的隐秘世界吗？

“有了足够的数据，数字就可以自己说话。”没门儿。

大数据的鼓吹者希望人们相信，在一行行的代码和庞大数据库的背后存在着有关人类行为模式的客观、普遍的洞察，不管是消费者的支出规律、犯罪或恐怖主义行动、健康习惯，还是雇员的生产效率。但是许多大数据的传道者不愿正视其不足。数字无法自己说话，而数据集——不管它们具有什么样的规模——仍然是人类设计的产物。

大数据的工具——例如 ApacheHadoop 软件框架——并不能使人们摆脱曲解、隔阂和错误的成见。当大数据试图反映人们所生活的社会化世界时，这些因素变得尤其重要，而人们却常常会傻乎乎地认为这些结果总是要比人为的意见来得客观些。偏见和盲区存在于大数据中，就像它们存在于个人的感觉和经验中一样。不过存在一种值得怀疑的信条，即认为数据总是越大越好，而相关性也等同于因果关系。

例如，社交媒体是大数据分析的一个普遍的信息源，那里无疑有许多信息可以挖掘。人们被告知，推特网的数据显示人们在离家越远的时候越快乐，而且在周四晚上最为沮丧。但是存在许多理由对这些数据的含义提出质疑。

首先，人们从皮尤研究中心获悉，美国上网的成年人中只有 16%使用推特网，因而他们绝对不是一个具有代表性的样本——与整体人口相比，他们中年轻人和城市人的比例偏多。

此外，人们知道许多推特账号是被称作“机器人”程序的自动程序、虚假账号或是“半机器人”系统（即得到机器人程序辅助的人为控制账号）。最近的估计显示，可能存在多达 2000 万个虚假账号。因此就算人们想要踏入有关如何评估推特网用户情绪的方法论雷场之前，请先问一下这些情绪究竟是来自真人，还是来自自动化算法系统。

“大数据将使我们的城市变得更加智能和高效。”在一定程度上是的。

大数据可以提供帮助改善城市的宝贵见识，但是它对人们的帮助仅此而已。因为数据在生成或采集的过程并不都是平等的，大数据集存在“信号问题”——即某些民众和社区被忽略或未得到充分代表，这被称为数据黑暗地带或阴影区域。因此大数据在城市规划中的应用在很大程度上取决于市政官员对数据及其局限性的了解。



例如，波士顿的 StreetBump 应用程序是一个比较聪明的以低成本收集信息的途径。该程序从开车经过路面坑洼处的驾驶员的智能手机上收集数据。更多类似的应用正在出现。但是如果城市开始依靠仅来自智能手机用户的信息，那么这些市民只是一个自我选择样本——它必然导致拥有较少智能手机用户的社区的数据缺失，这样的社区人群通常包括了年老和不那么富有的市民。

尽管波士顿的新城市机械办公室作出了多项努力来弥补这些潜在的数据缺陷，但不那么负责的公共官员可能会遗漏这些补救措施，最终会得到不均衡的数据，从而进一步加剧已有的社会不公。人们只要回顾一下曾经过高估计了年度流感发病率的 2012 年“谷歌流感趋势”，就可以认识到依赖有缺陷的大数据可能给公共服务及公共政策造成的影响。

在网上公开政府部门数据的“开放政府”计划——如 Data.gov 网站及“白宫开放政府计划”——也存在同样的情况。更多的数据未必会改善政府的任何功能，包括透明度和问责，除非存在可以使公众和公共机构保持接触的机制，更不用说促进政府解释数据并以足够的资源作出反应的能力。所有这些都非易事。事实上，人们身边还没有很多技能高超的数据科学家。各大学目前正在争相定义这一行当、制订教程和满足市场需求。

“大数据对不同的社会群体不会厚此薄彼。”几乎不是这样。

对大数据所号称的客观性的另一个期待是对于少数群体的歧视将会减少，因为原始数据总是不含社会偏见的，这使得分析可以在大规模的水平上进行，从而避免基于群体的歧视。然而，由于大数据能够作出有关群体不同行为方式的论断，它们的使用通常恰恰就是为了实现这个目的——即把不同的个体归入不同的群体中。例如，最近有一篇论文指科学家听任自己的种族偏见影响有关基因组的大数据研究。

大数据有可能被用来搞价格歧视，从而引发严重的民权担忧。这种做法在历史上曾被称为“划红线”。最近，剑桥大学对脸谱网 5.8 万个“喜欢”标注进行的大数据研究被用来预测用户极其敏感的个人信息，如性取向、种族、宗教和政治观点、性格特征、智力水平、快乐与否、成瘾药物使用、父母婚姻状况、年龄及性别等。

记者汤姆•福尔姆斯基这样评价该项研究：“此类容易获得的高度敏感信息可能会被雇主、房东、政府部门、教育机构及私营组织用来对个人实施歧视和惩罚。而人们没有任何抗争的手段。”



最后考虑一下在执法方面的影响。从华盛顿到特拉华州的纽卡斯尔县，警方正在求助于大数据的“预测性警事”模型，希望能够为悬案的侦破提供线索，甚至可以帮助预防未来的犯罪。

不过，让警方把工作专注于大数据所发现的特定“热点”，存在着强化警方对声誉不佳的社会群体的怀疑以及使差别化执法成为制度的危险。正如某位警察局长撰文指出的，尽管预测性警事算法系统不考虑种族和性别等因素，但是如果没有对差别化影响的考虑，使用这种系统的实际结果可能“会导致警方与社区关系恶化，让公众产生司法程序缺失的感觉，引发种族歧视指控，并使警方的合法性受到威胁。”

“大数据是匿名的，因此它不会侵犯我们的隐私。”大错特错。

尽管许多大数据的提供者尽力消除以人类为对象的数据集中的个体身份，但身份重新被确认的风险仍然很大。蜂窝电话数据看起来也许相当匿名，但是最近对欧洲 150 万手机用户的数据集进行的研究表明，只需要 4 项参照因素就足以挨个确认其中 95%的人员的身份。研究人员指出，人们在城市中走过的路径存在唯一性，而鉴于利用大量公共数据集可以推断很多信息，这使个人隐私成为“日益严重的担忧”。

但是大数据的隐私问题远远超出了常规的身份确认风险的范畴。目前被出售给分析公司的医疗数据有可能被用来追查到个人的身份。关于个性化医疗有很多谈论，人们的希望是将来可以针对个人研制药物和其他疗法，就好像这些药物和疗法是利用患者自己的 DNA 制作出来的。

就提高医学的功效而言，这是个美妙的前景，但这本质上依赖于分子和基因水平上的个人身份确认，这种信息一旦被不当使用或泄露就会带来很大的风险。尽管像 RunKeeper 和 Nike+等个人健康数据收集应用得到了迅速发展，但在实践中用大数据改善医疗服务仍然还只是一种愿望，而不是现实。

高度个人化的大数据集将成为黑客或泄露者觊觎的主要目标。维基揭密网一直处在近年几起最严重的大数据泄密事件的中心。正如从英国离岸金融业大规模数据泄露事件中看到的，与其他所有人一样，世界上最富有的 1%人口的个人信息也极易遭到公开。

“大数据是科学的未来。”部分正确，但它还需要一些成长。

大数据为科学提供了新的途径。人们只需看一下希格斯玻色子的发现，它是历史上最大规模网格计算项目的产物。在该项目中，欧洲核子研究中心利用 Hadoop 分布式文件系统对所



有数据进行管理。但是除非人们认识到并着手解决大数据在反映人类生活方面的某些内在不足，否则可能会依据错误的成见作出重大的公共政策和商业决定。

为了解决这个问题，数据科学家正在开始与社会科学家协作。随着时间的推移，这将意味着找到把大数据策略和小数据研究相结合的新途径。这将远远超越广告业或市场营销业采用的做法，如中心小组或 A/B 测试（即向用户展示两个版本的设计或结果，以确定哪一个版本的效果更好）。确切地说，新的混合式方法将会询问人们做某些事情的原因，而不只是统计某件事情发生的频率。这意味着在信息检索和机器学习之外，还将利用社会学分析和关于人种学的深刻认识。

技术企业很早就意识到社会科学家可以帮助它们更加深刻地认识人们与其产品发生关系的方式和原因，如施乐公司研究中心就曾聘请了具有开拓精神的人类学家露西•萨奇曼。下一阶段将是进一步丰富计算机科学家、统计学家及众多门类的社会科学家之间的协作——不仅是为了检验各自的研究成果，而且还要以更加严格的态度提出截然不同的各类问题。

考虑到每天有大量关于人们的信息——包括脸谱网点击情况、全球定位系统（GPS）数据、医疗处方和 Netflix 预订队列——被收集起来，人们迟早要决定把这样的信息托付给什么人，以及用它们来实现什么样的目的。人们无法回避这样的事实，即数据绝不是中立的，它很难保持匿名。但是人们可以利用跨越不同领域的专业知识，从而更好地辨别偏见、缺陷和成见，正视隐私和公正将面临的新挑战。

【原文链接】　　【回到目录】

安替：大数据时代的阶级斗争


安替：哈佛尼曼学者、专栏作家

“作为阶级斗争论的反对者、一个‘右派’，我认为大数据是中国和世界年轻人难得的成功自由通道，它体现的恶，必须要通过民主机制让恶与恶互斗以获得制衡，然后通过法治固定下来，最终保护普通民众（‘他们’）的权利。”

关于大数据会对社会产生的问题的讨论，我读过最深入的见地，莫过于克罗尔（Alistair Croll）2012 年 10 月的一篇博文：“可能数据驱动下的世界给人最大的威胁是道德方面的。我们的社会安全网由不确定性编织而成。我们有福利、保险等等机制，仅仅因为我们不知道未来会发生什么——所以我们以共享资源的方式分担风险。我们越是能预测未来，我们越不愿意和别人分享。”

但克罗尔在说这句话时，没有点名“我们”是谁。的确，有了大数据，可以预测一个人的购买习惯、健康状况、危机出现的地点和时间，其准确度也会随着技术的发展不断增进。虽不能就此断言这就是预测未来的工具，但有了它的确会拥有更准确的信息优势，足以获得巨大利益。很遗憾的是，大数据和互联网博客、微博等社交媒体的内容产生机制不一样，它并不是一个人人都能使用的工具，从诞生开始，它就更“亲”政府和企业。克罗尔能用“我们”来谈论大数据，也是因为他本人是互联网企业家。

最偏远地区的农民可以用几百元的国产手机上微博，但全国 3 亿用户的微博大数据只有新浪公司和中央政府网监部门才能染指，这成就了新浪微博的拆分上市，以及中央网监部门在政府内地位的提升，外加每年数百亿元与此相关的生意。在美国，情况也一样，最大的大数据处理机构是政府的国家安全局（NSA）；做未来危机预测花钱最多也做得最好的，是武器公司洛克希德•马丁（Lockheed Martin）和国防部的合作项目 ICEWS，而民间大数据翘楚如谷歌、Facebook 等企业，被迫让 NSA 在大数据上留有后门接口。

之所以产生这样的恶果，是因为个人建一个博客的时间成本大概是一小时、发一条微博的成本大概是一分钟，但大数据的收集、分析所需要的技术准备、存储资源、运用成本和编程维护不是个人所能承担的，所以个人从一开始就不能“拥有”大数据，而必须以某种方式“购买”其分析结果。这简直就是一个冤大头的年代：个人产生的社交信息被企业集中，经过分析，重新以各种方式卖给信息生产者本人。

拥有大数据资源和技术的企业，财富积累的速度是极其惊人的。目前正在全球各国蔓延的经济危机，并没有对这些人产生影响，无论在硅谷还是北京，每个月都有相关创投企业被购买、被追加数亿投资、不断产生亿万富翁的故事。

而能控制大数据公司的政府，无论是以秘密进行的方式（如美国），还是公开的方式（如中国），数据权力集中的速度也是惊人的。美国“9•11”之后建立的 NSA 大数据监控系统，如今已经到了漫天布网、可以预测潜在恐怖人士的地步；而中国公安部长郭声琨，也在 5 月 9 日明确指出要提高用大数据预防打击犯罪的能力。

也就是说，这个时代真正的“左派”，应当旗帜鲜明地反对这个从诞生起就透着资本和国家机器味道的新技术。因为从阶级斗争的角度，这个新技术，无论怎么玩，都无法拉平财富差距，只会快速积聚政府权力和企业资本。用阶级斗争的语言翻译克罗尔的话，就是在大数据时代，企业和政府（“我们”）越来越能预测公民（“他们”）的未来，并由此盈利和集权。

当然，我不是“左派”，不是阶级斗争理论爱好者。我相信资本和权力无法真正同流合污，他们之间的利益冲突，导致制衡出现，让整个社会走向改善。在谷歌、微软、雅虎等企业的逼迫下，美国白宫不得不在 5 月 1 日发表名为《大数据：抓住机会、保存价值》的白皮书，提出如何平衡大数据发展和公民权利之间关系的路线图。这也是历史上法治国家对人权的保护最终能驯服凶恶的资本主义的原因。

更重要的是，这项技术本身，并不是贵族技术，只要有相关程序员就可以。拥有大数据资源的网站公司本身，也完全可从一个小 App 开始。任何一个立志于网络创新事业的年轻人，只要视野和机会得当，都可搭上这轮风潮，依靠大数据成为克罗尔眼中“我们”的一员。在中国近年来因为社交媒体创新而致富的年轻人，也很少是“官二代”。在大数据技术造成的财富地壳位移的过程中，个人上升通道，至少对有心的年轻人来说，也算公平。

所以作为阶级斗争论的反对者、一个“右派”，我认为大数据是中国和世界年轻人难得的成功自由通道，它体现的恶，必须要通过民主机制让恶与恶互斗以获得制衡，然后通过法治固定下来，最终保护普通民众（“他们”）的权利。

（文章首发自财新网）

【原文链接】　　【回到目录】



林靖东：Fuzz：一个反“大数据”的流媒体公司

林靖东：腾讯科技作者

“

IBM 开发出来的沃森超级计算机将被用于处理成千上万的法律和医疗文件，以便为人们做出各种决策时提供支持。沃森很容易被推广应用到音乐、电影和书籍领域。然而，虽然这么做有利于提高销售业绩，但也有可能扼杀了文化的创新。沃森怎么可能预测出印象主义绘画、未来派诗歌或新浪潮电影的兴起呢？ 它怎么可能赞同斯特拉温斯基（Stravinsky）呢？ 而大数据也很可能会错过达达主义。

”

在 2013 创办的所有初创公司之中，Fuzz 肯定是最有诱惑力但却被大多数人忽视的一家公司。Fuzz 自称是一个“完全没有机器人因素的人力电台”，人们在发现新音乐方面越来越信任各种算法，Fuzz 对这一趋势提出了挑战。Fuzz 盛赞人工 DJ 发挥出来的重要作用，这里所说的人工 DJ 指的是 Fuzz 的一批固定用户，他们应邀将自己的音乐上传到 Fuzz 网站，以创建和共享他们自己的无线电台。

Fuzz 背后的想法或者说希望是，人工 DJ 可以传递出算法所不能传递的信息。它希望朝着与 Pandora 相反的方向发展，而后者主要是通过各种算法来完成所有繁重的工作。Fuzz 的创始人杰夫•亚苏达（Jeff Yasuda）去年 9 月在接受彭博社采访时称：“用户对这种助理型体验有着巨大的需求，我们只是想让大家相信，最令人信服的推荐只能来源于活生生的人。”

但是，虽然 Fuzz 的上线几乎没有引起任何人的关注，但是各种算法在艺术加工过程中所发挥出来的重要作用却变得越来越不容忽视了。最近，《沙龙》的技术批评家安德鲁•莱昂纳德（Andrew Leonard）在他撰写的一篇关于 Netflix 进军原创节目的服务——House of Cards 的评论文章中也强调了算法的重要作用。那些节目的原创性背后的秘密现在已经人人皆知了，即 Netflix 首先通过研究用户记录，发现重新制作一部同名的英国电视剧可能



会大获成功，尤其是让凯文•史派西（Kevin Spacey）来出演并由大卫•费恩切尔（David Fincher）来执导这部重新制作的电视剧。

莱昂纳德提出："在这个计算机算法已经成为最终的关注点的时代，这位著名导演还能继续生存吗？" Netflix 搜集了大量的用户数据，比如用户在观看某些电视剧的第一季节目时点击了多少次暂停按钮等等，他想弄清楚 Netflix 搜集的这些数据会对未来的电视剧造成什么样的影响。

很多其他的行业也面临着类似的问题。例如，亚马逊通过其 Kindle 电子书阅读器搜集了关于用户的阅读习惯的大量信息，包括用户们看完了哪些书？没看完哪些书？他们一般倾向于跳过哪些章节？哪些章节看得最为仔细？他们一般多长时间会去查一次字典以及在某些段落下面划上下划线？（其实并非只有亚马逊一家公司在这样做，其他的电子书阅读器也在搜集类似的数据。）

利用搜集到的这些数据，亚马逊就可以预测出能够让读者一口气读完一本书所需的所有元素。也许亚马逊甚至可以为读者提供其他的结局，能够令读者更欣喜的结局。正如最新发布的一份关于娱乐行业未来发展趋势的研究报告指出，我们现在所处的这个世界，很多事物都可以自行调整算法，以便建立一个更有魅力和互动性的未来。

获得了所有的用户数据后，Netflix 再不进入电影制作行业就太愚蠢了。正如 Netflix 一样，亚马逊也发现了这一点，因此它必须进入出版行业。然而，亚马逊的认识其实比 Netflix 还要深，因为它还经营着一个售书网站，它知道消费者所有的购买行为以及消费者愿意支付的价格是多少。现在，亚马逊经营着 6 种电子刊物，而且它还打算增加更多的刊物。

音乐行业在几年前就接受了类似的方法，搜集并建立了关于以前的热门歌曲和失败歌曲的庞大数据库，并借此来预测新歌曲是否可能成为热门歌曲。这种做法的优势是很明显的：新艺人无需拥有庞大的人脉关系也能与唱片公司签约，而在以前，人脉关系却是新艺人获得成功的必要条件之一。现在，新艺人只需利用过去的成功数据演绎出一首新歌，就很可能成为一首热门歌曲。

但是这种做法的劣势也很明显：我们最终能够听到的新歌可能听起来都差不多，缺乏创意和活力。正如克里斯托夫•斯泰因（Christopher Steiner）在他的新书《将此自动化》（Automate This）中所说："这样的技术也许会给我们带来新的艺人，但是由于他们的判

断完全建立于过去的流行歌曲基础之上，因此我们听到的新歌可能与我们已经忘记的那些流行小调都是同一类型的快餐作品。这显然是这种技术的弱点。”

IBM 开发出来的沃森超级计算机将被用于处理成千上万的法律和医疗文件，以便为人们做出各种决策时提供支持。由于需要阅读的文件太多，没有任何一位律师或是学会会员能够看完它们。如果目标只是分析过去出售过的商品并以此来预测未来可能畅销的商品，沃森很容易被推广应用到音乐、电影和书籍领域。

然而，虽然这么做有利于提高销售业绩，但也有可能扼杀了文化的创新。沃森怎么可能预测出印象主义绘画、未来派诗歌或新浪潮电影的兴起呢？它怎么可能赞同斯特拉温斯基（Stravinsky）呢？ 而大数据也很可能会错过达达主义。

想要了解算法给艺术创作造成的限制和提供的机会，我们就需要了解算法提供的机会通常是由 3 个要素组成，即发现、生产和推荐。象 Fuzz 那样的初创公司瞄准的是第三个元素即推荐，它寄望于某些用户希望由活生生的人而不是算法来为自己提供向导。

为喜爱读书的读者提供图书推荐服务的 Five Books 也实行了一种类似于 Fuzz 的模式，原因是它认为活生生的人在推荐图书方面会比死板的算法做得更好。亚马逊在图书推荐方面已经做得很不错了，但是 Five Books 以一种不同的衡量标准向读者推荐了保罗•克库格曼（Paul Krugman）、哈罗德•布鲁姆（Harold Bloom）和伊恩•麦克伊万（Ian McEwan）等作家的作品。其实在推荐方面，人工推荐和算法推荐是可以同时存在的，至少在可预见的未来是这样，因为读者们会在这两种推荐模式中找到一个平衡点。

但是在发现新人才和研究未来的创作方向时，算法的效果就没有这么乐观了。毕竟，只有确实存在伟大的作品时，算法推荐才有意义。如果算法选出的作品是建立在之前已经取得成功的作品和读者的及时反馈的基础之上，那么新作品的销量或许能够增加，但绝对不会为这个行业提供更有价值的好处。

最开始的迹象并不令人鼓舞。去年 12 月，英文版《环球时报》刊登了一篇关于本地朋克乐队熊战士（Bear Warrior）的报道，该乐队发现了一种巧妙的方法，能够检测出听众对他们的歌曲的反应。乐队主唱是北京大学精密仪器专业的一名研究生，他设计了一台名为“POGO 温度计”的设备，可以通过安装在音乐厅地毯中的一系列感应器检测出听众舞步的强度，然后将信号发送到一台中央计算机，最后让中央计算机对信号进行分析研究，帮助乐队改进他们的演绎方式。



据《环球时报》称，乐队发现，歌迷们会在鼓点敲响时开始摆动身体，而当主唱唱到歌曲的高音部分时，歌迷们跳舞的热情会达到顶点。正如乐队主唱所说："这些数据可以帮助我们了解到我们还可以如何去改善我们的演绎方式，让听众对我们的音乐作品作出我们希望看到的回应。"

或许，这确实有助于改善他们的演绎方式，但是朋克音乐什么时候变得这样细致入微了？让听众高兴是管理顾问需要考虑的事情，但绝不是朋克音乐人应该考虑甚至着迷的事情！性手枪乐队（Sex Pistols）唱歌时，音乐厅的地毯只有一个功能，那就是供歌迷们在上面跳舞，绝不会安装什么感应器。但是性感手枪乐队却创造了朋克音乐的一场革命，而熊乐队，顶多只能将朋克音乐变成他们谋生的职业。

【原文链接】 【回到目录】

康国卿，柳小龙：反思大数据时代：一种全景敞式监狱的视角

康国卿：江西师范大学传播学院新闻学硕士研究生
柳小龙，浙江工业大学人文学院新闻传播学硕士研究生

“

大多数的网民在互联网中以无意识的状态生产内容，丝毫没有注意到‘第三只眼’时时刻刻在盯着自己、跟踪自己。‘我们时刻都暴露在“第三只眼”之下：亚马逊监视着我们的购物习惯，谷歌监视着我们的网页浏览习惯，而微博似乎什么都知道，不仅窃听到了我们心中的‘TA’，还有我们的社交关系网。’伴随着物联网的发展，不仅仅是互联网，甚至整个社会俨然成了一个全景敞式监狱。

”

摘 要：近两年来“大数据”成为业界和学术界舌尖上的热词，在对“大数据”这一话题的讨论中涉及其应用方面的较多，而对其进行批判性反思的则相对较少。福柯的全景敞式监狱理论却为喧嚣的大数据讨论打了一剂镇定剂。福柯的监视概念在大数据时代背景下出现了数字化、实时性、隐蔽性和预测性四个新特点。大数据时代下的监视作为一种手段，其最终目的是通过用户数据进行更加精准的市场定位。其对市场效益的追求大于其对信息利用的民主追求。对此从法律制度和职业伦理这两个角度构建用户隐私保护防线具有一定的借鉴意义。

关键词：大数据；全景敞式监狱；监视；消费社会；隐私保护

一、引言

对于大数据（Big data）的定义，简单来讲就是，“无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合。[1]” 它主要有四个特征，即四个“V”：Volume（容量），Variety（种类），Velocity（速度）和最重要的 Value（价值）。“Volume 是指大数据巨大的数据量与数据完整性。[2]” Variety 主要指大数据的种类繁



多，尤其是 Web2.0 时代的到来以及在以互联网为应用平台的支持下，用户自己生产内容（UGC），其内容呈现的多样性也就不言而喻了。Velocity 主要指出的是数据生成的实时性以及接收的实时性。Value 指大数据时代下，显现的数据虽然是半结构化或者非结构化的，但经过数据的结构化处理以后，散如砂砾的数据却可以显现出巨大的价值。

虽然大数据时代的到来给我们的生活、工作、学习带来了诸多便利，但任何事情都存在两面性，大数据时代也意味着用户的很多信息暴露于各种数据大亨、数据分析公司的眼皮底下，受到其监控。本文主要着眼于福柯的全景敞式监狱的视角，剖析身处大数据时代下的我们，类似于处在全景敞式监狱的环境当中。数据大亨拥有信息数据处理的绝对技术手段，处处并且实时监视着我们的行为，刺激我们的消费欲望，并诱导我们形成符号消费、超前消费、消费至上的消费主义文化观念。

二、大数据的监视——黑暗中的“第三只眼”

大多数的网民在互联网中以无意识的状态生产内容，丝毫没有注意到“第三只眼”时时刻刻在盯着自己、跟踪自己。“我们时刻都暴露在‘第三只眼’之下：亚马逊监视着我们的购物习惯，谷歌监视着我们的网页浏览习惯，而微博似乎什么都知道，不仅窃听到了我们心中的‘TA’，还有我们的社交关系网。[3]” 伴随着物联网的发展，不仅仅是互联网，甚至整个社会俨然成了一个全景敞式监狱。当然，大数据时代下的全景敞式监狱，不论是监视者还是被监视者，以及监视方式等都已发生了极大的变化，但其原理还基本相同。

全景敞式监狱（panopticon）最早是由功利主义哲学家边沁（Jeremy Bentham）设计，其设计结构为“四周一个环形建筑，中心是一座瞪望塔。瞭望塔有一座大窗户，对着环形建筑。环形建筑被分成许多小囚室，每个囚室都贯穿建筑物的横切面。各囚室都有两个窗户，一个对着里面，与塔的窗户相对，另一个对着外面，能使光亮从囚室的一端照到另一端。然后，所需要做的就是在中心镣望塔安排一名监督者，在每个囚室里关进一个疯人或一个病人、一个罪犯、一个工人、一个学生。通过逆光效果，人们可以从瞭望塔的光源恰好相反的角度，观察四周囚室里被囚禁者的小人影。这些囚室就像是许多小笼子、小舞台。[4]” 随着互联网的发展，这种结构成为了大数据时代的一种建筑学形象，且监视出现了一些新的特点。

（一）监视的数字化

监视的数字化主要体现在对用户行为记录的数字化上，其以数字化记忆的方式得以实现。“数字化记忆作为一种全景控制的有效机制，不仅支持了对等级森严的机构和社会的控



制，并且还会去寻找对他们自身的支持，从而巩固并加深现有的（不平等）信息权力分配。[5]” 尼葛洛庞帝（Nicholas Negroponte）早在 1996 年就提出了“数字化生存（Being Digital）”的概念，他指出人类生存于一个虚拟的、数字化的活动空间，在这个空间里人们应用数字技术从事信息传播、交流、学习、工作等活动从而构成了数字化的生活方式。数字化技术是数字计算机一切运算和功能的基础，0 和 1 是各种信息最基本的数字化表达，而它们的组合却引发了广泛的数字化革命，电视、广播、电影、通信网络都已步入了数字化的发展进程。

数字化让我们接受到越来越多的信息，体验到越来越多的乐趣，而互联网在提供给我们极大便捷的同时也在发生着深远的变化。在 Web1.0 时代互联网仅仅是信息访问、浏览的工具，但是在 Web2.0 时代的今天，互联网还是信息共享的工具。人们的数字化活动正在不断地被记录到数字储存器中。从使用电子邮箱到浏览购物网站，一举一动都被一只无形的眼睛监视着。这些呈爆炸式增长的数字信息不断汇集，人们的数字化行为数据得到充分地量化，即便是剥离情景的片段信息，网络平台也可以利用计算机对其进行准确预测和重组。在获得看似更为人性化的服务之外，个人数据是否被少数人应用来获取大多数人的隐私，和控制大多数人利益的问题已成为矛盾的焦点。

（二）监视的实时性

监视的实时性主要指大数据时代背景下，我们的行为随时随地受到监视。数据化的文字输入是现代生活必不可少的日常行为，在电脑以及手机上工作、交流、娱乐都离不开高频率的文字输入。看似微不足道的动作，其背后却隐藏着人们生活习惯、行为取向的综合记录，这些数据在大数据时代有着举足轻重的价值。输入法软件能为我们回想印象模糊的词句，搜索网站能为我们提供最想知道的答案，电子商城能为我们提供我们最想购买的商品，微博的定位功能可以随时随地显示我们所处的位置。通过互联网，我们的想法似乎时刻被他人洞悉。

对文字输入信息的收集，可以看作是一种日常生活中不间断的信息监视，而数字化记忆又形成了一种更为严酷的数字圆形监狱。伴随着每一次敲击电脑键盘，触摸手机屏幕的文字输入，我们的所想所为被时时刻刻地记录下来，这些信息再被他人经过数据化的处理实现影响实施的最大化。

（三）监视的隐蔽性

“‘可见性’是圆形监狱构成并正常运作的一个关键要素。被监视者和监视者，一个在明处，一个在暗处。于是，在明处的囚犯随时能被监视却看不见监视者，而在暗处的监视者能清楚地观察到在明处的囚犯的一举一动，却不被发现。[6]” 在传统时代的圆形监狱中，被监视者虽然看不见监视者，但可以知道自己受监视这一事实。然而大数据背景下，圆形监狱出现的一个明显的新特点是，用户在应用互联网之时很多时候全然不知背后有“一双眼睛”一直在盯着自己。在被监视者心中，监视者以缺场的形式存在。这大大减少了用户对被监视的忌惮心理，互联网应用的匿名性更使其对自身的网络行为降低警惕，遗留下更多的蛛丝马迹，这为监视者分析用户行为提供了大量的非结构性数据。

（四）监视的预测性

在以往全景敞式主义的视角下，监视者可以对被监视者的言语、行为进行记录并加以保存，但对被监视者未来将要做出的行为却难以预测，这种预测难以实现的一个重要的原因是被监视者忌惮于自身所处的被监视的环境，他们会对其行为加以自我约束，可能不会完全表现出真实的自我。然而身处大数据时代下的被监视者则不一样，基于监视的隐蔽性和网络的匿名性心理，用户会更加真实的表露自身的言语和行为。这为监视的预测性提供了更真实的信息。“大数据的核心就是预测”，“大数据不是要教机器像人一样思考”。

相反，它是把数学算法运用到海量的数据上来预测事情发生的可能性。[7]” 明显的例子就是，当当网会根据用户在其网站浏览、收藏以及购买的图书遗留的痕迹，不定时的向用户推荐同类别的图书，搜狗输入法会根据你平时打字的习惯，把你平时常用的字词以及短语放在最前面，人人网等社交网站会根据你平时浏览、发表以及注册的信息不时地向你推荐你的好友动态等。随着云计算的进一步发展，监视者对大数据收集、分析的能力会越来越强，相应的对被监视者行为的预测也会越来越精准。

三、大数据的诱导——消费的“揠苗助长”

社交网络的发展、计算机技术的进步为大数据的挖掘提供了更加便捷的条件。大数据时代下监视只是监视者收集、记录用户信息的一种手段。利用数据挖掘技术分析用户的行为习惯、特点、喜好，预测其消费行为，最终向其推送消费产品信息，促使其进行消费，甚至引导用户的消费观才是最终目的。大数据的挖掘强化了消费至上的观念成为现今消费主义不断盛行的帮凶。鲍德里亚在其著作《消费社会》中指出“我们今天到处被消费和物质丰富的景象所包围，这是由实物、服务和商品的大量生产所造成的。”大数据挖掘催生出的精准化营销，以推送更适合用户口味的消费信息刺激用户进行消费，构筑消费主义文化。



在这一过程中主要基于用技术手段对消费行为进行实时监测，以及对消费行为的数据分析催化消费主义观念这两个方面。

（一）被技术统治的消费行为

大数据四个特性中最后一个是 Value（价值），其实现途径在于对浩瀚的半结构化、非结构化数据的收集和分析。如马克•波斯特（Mark Poster）在《第二媒介时代》中所言，“在经济领域内，进行网上销售的零售商把他们积累的客户资料视为自己的财产，视为一种有价资产，这种资料是他们从销售中获得的副产品，他们还可以进而卖给其他零售商。[8]” 监视者根据用户的分析报告，针对不同类型的用户适时推出适宜的产品信息。

近几年逐渐兴起的“精准广告”就属于这一应用的典范，指的是“在对受众行为进行准确研究的基础上，借助现代信息技术手段和强大的受众数据库资源而建立的个性化的受众沟通服务体系，在实现企业广告精准投放目标消费者的同时，也给消费者创造了满意的沟通体验。[9]” 其技术支持主要源自云计算的发展，“云计算通过数据仓库，将分散的海量数据（PC 端、手机端、应用端等）统一导入云端平台，在大规模分布式计算机群上实现数据的整合、管理、挖掘、建模和应用。[10]” 并且数据监视者“根据用户上网的综合行为来分析他的特征，包括用户注册时的一些基本信息：他搜索过什么广告，浏览过什么样的网页以及在页面的停留时间等。通过对这些信息的提取和分析可以准确识别用户特征获得用户消费需求，从而锁定目标用户进行精准广告投放。[11]”

（二）被催化的消费主义观念

在数字圆形监狱下，数据大亨掌握用户大量的数据，并依此作为用户行为研究、刺激用户消费欲望的基础。这一现象固然要引起我们的警惕，但是，最令人担忧的莫过于，根据分析结果不断对用户进行消费欲刺激催生消费至上的消费观念。

首先是引导用户进行符号消费。传统消费者主要关注的是商品的使用价值，而现代消费文化崇尚的是符号的消费，引导消费者满足于其对商品建构的意义。“人们购买某种商品和服务，主要不是为了它的实用价值，而是为了寻找某种感觉，体验某种意境，追求某种意义。而这种符号的生产和营销，意义的生产和营销是绝对离不开媒体的参与，离不开现代传媒的推波助澜的。[12]” 大数据公司通过对用户数据的挖掘，了解用户心理，将各种符号化了的商品通过互联网应用平台不断地推向用户的眼前，促使其进行消费。除了符号化了的商品之外，与此相伴随的还有劝服性的话语，其主要是以符号的“示同”和“示异”功能得以体现。“所谓‘示同’，就是通过消费来表现与自己所认同的某个社会阶层的相



同、一致和统一。所谓‘示异’，就是通过消费显示与其他社会阶层的不同、差别和距离。两者结合就使消费者获得了在社会中的自我定位。[13]” 如此一来，掌握着用户行为动态的数据大鄂，根据用户的自我认同心理，不断地诱导其进行符号消费。

其次是刺激用户进行超前消费。在大数据时代背景之下，众多的电子商务公司掌握用户大量的购物数据，在不断刺激用户消费欲望的同时，难免会塑造用户的超前消费观念，最终导致用户形成习惯性超前消费行为。根据中国互联网网络信息中心 2013 年 3 月份发布的《2012 年中国网络购物市场研究报告》显示：2012 年，我国网购用户人均年网购消费金额达到 5203 元，与 2011 年相比增加 1302 元，增长 25%。2012 年，用户网购频次有了显著的提升，用户半年平均网购次数达到 18 次，较 2011 年增加 3.5 次。同时，半年网购 10 次以上的用户比例提升较快，增加了 23.8 个百分点，达到 54.5%。这些数据虽然未能直接反映用户的超前消费行为，但从用户人均年网购消费金额和网购频次的增幅来看，一些用户的超前消费行为以及潜在的超前消费观念也难以被抹去。而背后的一大推手则少不了电子商务公司基于用户数据的云计算分析。

四、结 语

不可否认大数据时代的到来，为我们的学习、生活带来诸多便利。然而这并不代表大数据时代的到来可以解决一切难关，其试图突破的道德枷锁，引领的消费倾向等问题仍未得到妥善的处理。甚至蔓延到网络信息领域，如最近美“棱镜”监听项目的曝光，层出不穷的企业泄露客户资料事件，又涉及网络安全问题，再次向我们发出大数据时代下个人隐私保护的预警。

对此立法是最好的解决途径，完善的立法不仅对保护用户隐私来说极其重要，而且对政治民主进程的推进也比不可少。例如“规定只有用户需要个性化服务定制的时候，提出需求，大数据公司才可以调用该用户的信息，其他情况下的信息调用都采取匿名的方式，否则就视作侵犯隐私。[14]” 同时这也是对用户的一种尊重。立法固然能在机制层面解决问题，然而落实到微观层面还是需要相关从业者加强自身职业道德修养，提高自身的约束力，在未获得允许的情况下坚决不越“雷池”半步。“如果不从法律制度和职业伦理双重角度去构建保护个人网上隐私的防线，任由各种利益集团和个人无限制地挖掘和疯抢，乃至不断转手倒卖，那么我们每一个人都将丢掉自己最后一块‘遮羞布’。我们将重新回到蒙昧时期的‘丛林’，再一次‘重新部落化’成为原始人。[15]”



任何技术决定论、工具理性或者过度乐观主义都只能是问题认识的一个方面。正如伽达默尔（Hans-Georg Gadamer）所言，任何历史都是“效果历史”，是偏见组成了我们的认识。对大数据多角度认识同样也是如此，多维度的分析可以为我们提供重要的参考价值。本文基于福柯的“全景敞式监狱”视角来审视大数据时代的到来，希望给一路昂扬的“大数据”赞歌带来一次理性的降温。

作者简介：康国卿，男，江西师范大学传播学院新闻学硕士研究生；主要从事新闻实务研究；柳小龙，男，浙江工业大学人文学院新闻传播学硕士研究生，主要从事传播学研究。

参考文献：

[1] 张意轩,于洋.大数据时代的大媒体[N].人民日报,2013-01-17.

[2] 陈昌凤.“大数据”时代如何做新闻?[J].新闻与写作,2013(1).

[3] 余建斌,赵展慧.大数据崛起[N].人民日报,2013-02-22.

[4] (英)维克托?迈尔-舍恩伯格等著.盛杨燕等译.大数据时代[M].杭州:浙江人民出版社,2013.

[5] (英)米歇尔•福柯著.刘北成等译.规训与惩罚[M].北京:三联书店,2007.

[6] (英)维克托•迈尔-舍恩伯格著,袁杰译.删除:大数据取舍之道[M].杭州:浙江人民出版,2013.

[7] 梁婷.“圆形监狱”的隐喻[D].西南政法大学,2006.

[8] (美)马克•波斯特著.范静哗译.第二媒介时代[M].南京:南京大学出版社,2000.

[9] 黄薇.3G 时代手机精准广告传播策略[J].新闻爱好者,2010(7).

[10] 范晓东.好耶:精准广告不是谎言[J].互联网周刊,2012(7).

[11] 李娜,李爱军.基于用户特征分类的精准广告投放研究[J].电脑知识与技术,2010(1).

[12] 都错了！阿里要用新浪微博的数据干这[EB/OL].http://www.huxiu.com/article/13911/1.html,2013-05-02.

[13] 许永.现代传播与消费文化[J].新闻知识,2002(12).

[14] 许薇.论大众文化背景下的消费文化[D].浙江师范大学,2006.

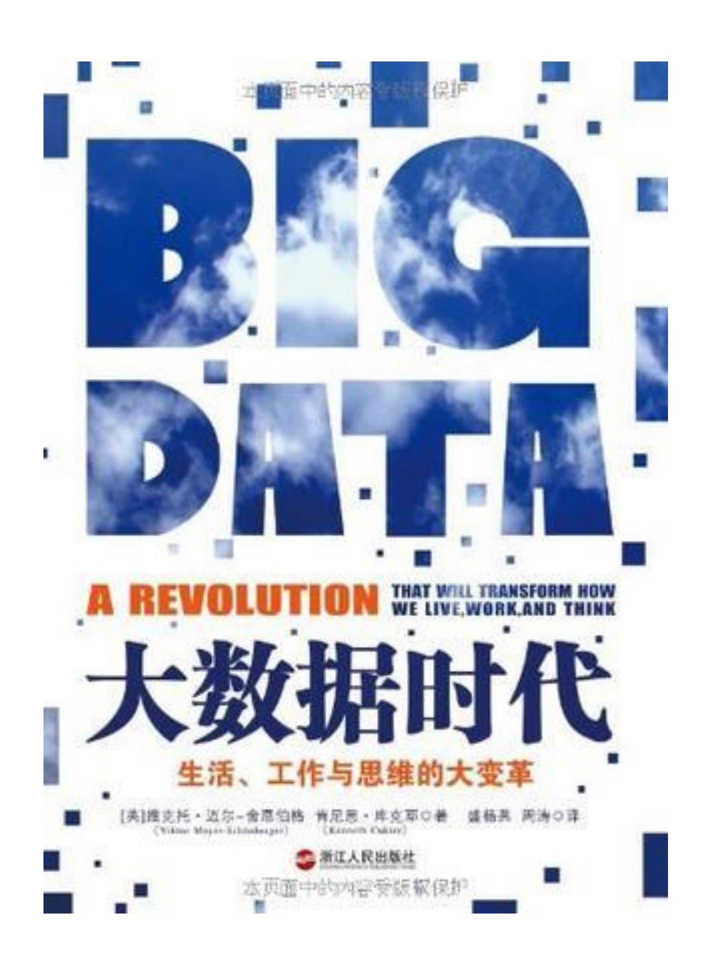
[15] (印度)艾可里•桑多萨姆,大数据时代:隐私保护靠立法[J].中国经济周刊,2013(31).

[16] 陆高峰.大数据时代的公众隐私[J].青年记者(上),2013(7).

【原文链接】 【回到目录】



推荐阅读：《大数据时代》



作者：[英] 维克托•迈尔•舍恩伯格（Viktor Mayer-Schönberger）

出版社：浙江人民出版社

副标题：生活、工作与思维的大变革

原作名：Big Data: A Revolution That Will Transform How We Live, Work, and Think

译者：周涛

出版年：2012-12



页数：261

定价：49.90 元

《大数据时代》是国外大数据研究的先河之作，本书作者维克托•迈尔•舍恩伯格在书中前瞻性地指出，大数据带来的信息风暴正在变革我们的生活、工作和思维，大数据开启了一次重大的时代转型，并用三个部分讲述了大数据时代的思维变革、商业变革和管理变革。维克托最具洞见之处在于，他明确指出，大数据时代最大的转变就是，放弃对因果关系的渴求，而取而代之关注相关关系。也就是说只要知道“是什么”，而不需要知道“为什么”。这就颠覆了千百年来人类的思维惯例，对人类的认知和与世界交流的方式提出了全新的挑战。

【原文链接】　　【回到目录】



【换个角度玩.大数据】

侯世达：关于思考，我一直在思考



侯世达：美国学者、作家。主要研究领域包括意识、类比、艺术创造、文学翻译以及数学和物理学探索。《哥德尔、埃舍尔、巴赫：集异璧之大成》作者

“

侯世达认为，思考的关键在于美、在于品味，与逻辑或真理无关。这与大数据、大算法的现代人工智能格格不入。‘形式化的研究方法得出的是极其生硬的“智能”，毫无洞见可言。’ 他的人生目标是创造出许多绝美的事物。他选择了一条少有人走的研究道路，他在路上遇到了许许多多至臻至美的事物，他说‘我宁愿当个独立思考的人，不总是站在人们注意力的最前端。我觉得不被大多数人注意到没什么不好；但我相信最终我的想法会被更多的人知道。’

”

编辑的话：“侯世达”是 Douglas Hofstadter 的中文名，这个 1997 年由他的中文出版商所定的名字，如今已是他在中文世界里的通称，这个名字也确实比他的父亲、1961 年诺贝尔物理学奖得主、物理学家罗伯特•霍夫施塔特（Robert Hofstadter）按照姓名音译规则对应过来的中文名要好听。不过，侯世达还有一个更私密、也更漂亮的中文名，那就是 1976 年他的第一位中文老师高先生为他取的“侯道仁”。

与他的中文名字同样精彩的，是侯世达的成名作“Gödel, Escher, Bach: an Eternal Golden Braid”的译名——《哥德尔、埃舍尔、巴赫：集异璧之大成》。侯世达的这本书在英文世界里被简称为“GEB”——取哥德尔（Gödel）、埃舍尔（Escher）、巴赫（Bach）的首字母，而中文则以“集异璧”应对。



《集异璧》探讨了庞杂的主题，正如侯世达本人在该书出版 20 周年的再版前言中所写："……包括赋格和卡农，逻辑和真理，几何学、递归、句法结构、意义的本质、佛教禅宗、悖论、脑和意识、还原论与整体论、蚂蚁群落、概念和心理表征、翻译、计算机和计算机语言、DNA、蛋白质、基因编码、人工智能、创造性、意识和自由意志——偶尔还写到了音乐和艺术，它写到了所有的一切！很多人觉得不可能找到这本书的重点。"

但这本书还是有重点的。总体上说，《集异璧》被归为人工智能的经典著作，就像研究认知科学、心智哲学（Philosophy of Mind）、计算机科学、心理学、比较文学和物理学的侯世达被视为人工智能领域不可忽视的代表一样。上世纪 70 年代，侯世达痴迷于"思考是什么？"，投身于其时刚刚兴起的人工智能领域，他在《集异璧》中对计算机、程序、思考和大脑的描绘，开启了整整一代年轻人对 AI 的探索。但是，在人工智能领域掀起一个高潮后，侯世达却从公众的视野中消失了。

原因很简单：算法很巧妙、也能完成不少实际任务，但依托这种思路做出来的计算机并没有真正在"思考"。意识到这一点，侯世达对普通的人工智能彻底失去了兴趣，他自己的研究也转而建立在跟常规 AI 完全不同的技术上面。侯世达在美国印第安纳大学的研究小组叫做灵活类比研究小组（Fluid Analogies Research Group，FARG），"在 FARG 我们没有致力于开发实际的应用，诸如翻译引擎、答问机器、网络搜索软件此类的东西。我们只是在努力地理解人类概念的本质和人类思考的根本机制。我们更像是哲学家或试图探究人类心智奥秘的心理学家，而非旨在制造聪明的计算机或机灵程序的工程师。我们是一群老派的纯粹主义者，我们的动力源于内心深处的哲学好奇心，而不是制造实用设备的欲望（遑论赢得大笔金钱的欲望！）。"

这些年来，关于"思考是什么"，侯世达取得了一定进展，但更多的还是失败——FARG 开发的程序常常得出可笑的结果，远远谈不上"智能"。不过，侯世达看着这些失败"很开心"，因为"要是我们的任何系统真的在其微领域中获得了与人类相颉颃的智力，我们将痛心至极，因为那将是很可怕的：这意味着人的智力并非如我们所想的那样复杂或深奥。这意味着短短几十年的研究就足够人类解开人类思维的奥秘"。在他看来，程序真正具有智能将是人类的悲剧。

侯世达认为，思考的关键在于美、在于品味，与逻辑或真理无关。这与大数据、大算法的现代人工智能格格不入。"形式化的研究方法得出的是极其生硬的'智能'，毫无洞见可言。" 他的人生目标是创造出许多绝美的事物。他选择了一条少有人走的研究道路，他在路上遇到了许许多多至臻至美的事物，他说"我宁愿当个独立思考的人，不总是站在人们

注意力的最前端。我觉得不被大多数人注意到没什么不好；但我相信最终我的想法会被更多的人知道。”

“至于有没有可能我选错了路，这当然是可能的，但我并不担心这一点。人生苦短，我相信我自己的观点，而且我会捍卫它们。毕竟，俗话说得好，你都不相信自己，谁还会呢？”

果壳网带你走近侯世达，谈谈他心目中的思考、大数据、美，还有人生。

上世纪 70 年代，侯世达凭《哥德尔、埃舍尔、巴赫：集异璧之大成》开启了整整一代年轻人对 A.I.的探索。但是，在人工智能领域掀起一个高潮后，他却从公众视野中消失了，因为他认为，计算机并没有真正在“思考”。

果壳网：这么多年来你一直在思考“思考是什么”，你得出什么结论了吗？

侯世达：嗯，关于思考究竟是什么，我当然有所思考——很多的思考。我无法告诉你我在这个大问题上的全部思考。不过，就这次采访而言，让我姑且这么说：思考就是尝试去触及你身处境况的本质。我解释一下这句话的意思。

每当我说话或写作时，我总是在寻找用最恰当的字词传达我的意思。我每每惊讶于多少次我话说到一半停下来问我自己和我的听众，“我要说的那个字是什么来着？”我真的对此很在意，会尽可能地去找到它、也很感激对方的建议。有时候在共同努力下我们找到了我想要的表达，有时候则找不到。但一旦找到了，就是莫大的欢喜和畅快；找不到时，我总是难以释怀。对我而言，在所有的表达中找到最精准的字词非常之重要，因为只有它们才真正切中了我要说的意思。

我的夫人林葆芬是中国人，她常常会无意识地使用形象的四字成语为她的中文语句添色。唉，这些成语我是几乎听不懂，但换做母语是中文的人，不但会认得，还会觉得这是简洁生动表词达意的好法子。比方说，假如葆芬看到有人在一场精彩的舞蹈演出中哈欠连天，她或许会说这是对牛弹琴。这一形象的中文短语可谓“敲钉子敲中了钉子头”（用一个美国的俗语，意思是“正正好抓住了本质”）。要是我用英语描述这个烦人的哈欠家伙，我大概会说这是把珍珠洒在了猪面前（casting pearls before swine）。而我若是用法语，我会说“还不如拿果酱去喂猪”（autant donner de la confiture aux cochons）。这三种表达都既生动又形象，每一种在其母语讲述者看来都抓住了情况的本质。能够用简练而



熟悉的短语囊括复杂情形的关键所在是件非常开心的事情。这就像是做了一次绝妙而又精确的简化——漂亮地将榔头不偏不倚敲在了钉子头上。

如果我去商店不止买了一瓶我要的牛奶，还捎上了一盒我儿子想要的麦片，我可以用“一块石头打死了两只鸟”（I killed two birds with one stone）来概括这次小小的双重收获，指我出去一趟达成了两个目的。在中文里，你也许会说“一箭双雕”。这两个词组都简明扼要地将我“双重行为”的显著特征表达了出来，而说话者——在这里就是我自己——则会当即体味到求仁得仁的喜悦体验。要是我用中文说出来，我会比中文母语者还要喜出望外，因为一个词恰逢其时跃入脑中，这种感觉是名副其实（哪怕微乎其微）的胜利。

再来看另一种截然不同的情形。假设一位紧张的主人请了 20 位客人吃大餐。轮到甜点时，他决定上一种非常高档的冰淇淋，然后，在最后一刻他突发奇想在每份冰淇淋上都撒了辣椒粉。这个举动显得非常荒谬，因为它破坏了原本已经完美的东西。像我这样的人，都会希望找到最好（即最简洁、最鲜活）的方式概括他对这位主人考虑欠佳的行为的看法。在我看来中文里恰好有一个描述这类情况的成语。你能想到这个词吗？你是没费力气一下子就想到了，还是得有意识地找一会儿？

我认为，说出这样一个词的喜悦和一个人对这个词的理解深度成正比。字词越是能准确地概括你的想法，找到并说出它的喜悦也就越大——尤其是在你能迅速做到这一点的情况下。速度绝对很重要，因为想到得太晚就没用了。在法语里有个短语专门形容事后才想到该说的话那种强烈的挫败感：“l’esprit de l’escalier”，直译过来就是“楼梯妙语”。这个短语描绘的场景是：聚会上有人对你出言不逊，但你在种种社会压力之下，没能做出有力的反击。你觉得输了。稍后，离开聚会后不久，在你下楼梯的时候，一句完美的回应划过你的脑海——但现在为时已晚。找到了朝刚才那个讨厌的土老帽儿抡过去的绝好棒子却无处可施，其懊悔之情可想而知。这就是“楼梯妙语”的概念，这个短语凸显了迅速找到当前情况核心的重要性。现在你知道了一个表达特定社交场合本质的法语短语（至少知道了它的英语版），它会适时从你脑子里冒出来。新添了这个小小的“思考工具”，你的思想又充实了一点。

刚才你想到了形容那位慌张主人傻决定的“那个”四字成语了吗？很可能你想到了，但万一没有的话，答案是：画蛇添足。我希望你同意我说的这个短语将选择在冰淇淋上洒辣椒粉的精髓表达了出来，而找到这一精准凝练的概括——这个小“金块”——是很开心的事情。说到金子，在英语里我们会说那位慌张的主人到头来“给百合花镀金”（意思是，将一种美丽的花儿染成金色以作修饰，也因此将其美破坏殆尽）。

所有上述例子都显示了人脑是如何受驱使去尝试总结当前所处的境况——找到其本质、确定其关键概念。你的思维无时无刻不在试着将你所处的新情境和你曾经遇到过的其他情况作比较，也即在更高的抽象层面上理解新的情境；这涉及剔除无关和多余的细节，分离出真正重要的东西。如果你在各种情况中都能很好地做到这一点，意味着你是一个善于思考的人。

简言之，思考包含了刨去表面看深层。

【原文链接】　　【回到目录】

主编：方可成

编辑：王陶陶

设计：潘雯怡，季文仪

校订：童亚琦

出品人：杜婷

若希望订阅此电子周刊 doc 版请发空邮件至 cochinaweeklydoc+subscribe@googlegroups.com；

若订阅 pdf 版请发送至 cochinaweeklypdf+subscribe@googlegroups.com；

若订阅 mobi 版至 cochinaweeklymobi+subscribe@googlegroups.com；

若订阅 epub 版至 cochinaweeklyepub+subscribe@googlegroups.com。

此电子周刊由「我在中国」（Co-China）论坛志愿者团队制作，「我在中国」（Co-China）论坛是在香港注册的非牟利团体，论坛理事杜婷、梁文道、闾丘露薇、周保松。除了一五一十周刊之外，Co-China 每月还在香港举办论坛，并透过网络进行视频、音频和文字直播。2012 年开始 Co-China 在香港举办面向青年的夏令营，第一届主题为「知识青年，公共参与」，2013 年夏令营的主题是「始于本土：本土、国家、世界冲撞与协商」。

Co-China 论坛网址：https://cochina.co

Co-China 论坛新浪微博：CoChina 論壇（http://weibo.com/1510weekly）

Co-China 论坛 facebook：「我在中國」（Co-China）論壇
（https://www.facebook.com/CoChinaOnline）

